Endorsements

As more layers are added to our Broadband "onion", the second edition of Modern Cable Television Technology has replaced the first edition copy I have permanently positioned on the corner of my desk. *Modern Cable Television Technology* is to the cable industry, what the *Physicians Desk Reference* or PDR is to the medical profession, and every person involved in the industry should have a personal copy of this invaluable reference. It continues to be the most complete and up to the date technical reference book available.
—Steve Allen, BCE, BTCS, Sr. Broadband Technologist, Kramer. Firm, Inc.

Wow! What an outstanding contribution to the entire telecommunications industry this book has become! It's an absolute "must have" for any serious engineering professional in this or related industries. I've personally known and respected all of the authors for many years, and with the publication of this book, once again, they have proven their preeminence in the telecommunications engineering profession. Thanks guys, for a truly excellent reference book!
—Chris Bowick, Senior VP of Engineering and Chief Technical Officer, Cox Communications

This is "The Book" on cable technology. As expected from this group of authors, this text is thorough, accurate, and easy to read.
—Tom Elliot, Principal, T G E & Associates

Modern Cable Television Technology is written by some of the greatest technical minds in our industry. It is a terrific reference book for all cable engineers and technologists to have on their bookshelf and a wonderful educational tool for all technical trainers in the industry.
—Marwan Fawaz, CTO, Adelphia Communications

This book is required reading for engineers and other technical readers interested in a career in telecommunications, which in the next millennium will rely heavily upon a cable television infrastructure. Cable technology has emerged as a leading architecture for delivery of IP broadband data. This book offers the best comprehensive look at cable television technology that I have seen. I congratulate the authors on a job well done.
—Richard Green, President and CEO, CableLabs

Modern Cable Television Technology 2nd Edition is the best treatise yet on cable telecommunications technology. In my view it is a "must" for every cable television and broadband engineer or technologist to have by their desk. I've very frequently used the 1st Edition over the last 4 years both to refresh myself on various aspects of cable telecommunications as well as to direct others to specific paragraphs or sections when they have been seeking clarification about this or that aspect of "broadband". This 2nd Edition, like the 1st is easy reading even for the non-technical audience. In fact I've frequently recommended to analysts from the financial sector to obtain a copy for their reference as they study the Cable Industry. The authors have done a tremendous job in bringing this 2nd Edition right up to date.
—Nick Hamilton-Piercy PEng., Senior Technology Advisor, Rogers Cable Inc.

I have certain books in my library that are special, they are so concisely written that they always answer a question or promote clear thinking. These books, from another era, clearly explain radio physics and television; they are now joined by *Modern Cable Television Technology* as it too is a classic work on the art.
—Ted Hartson, Principal, Scottsdale Television Labs.

No question that this book gets referred to more than any other in my collection. The authors wrote what I consider to be the absolute best cable engineering reference book. Period. I've got a ton of small sticky notes marking various pages. If I could have only one cable-related engineering book, this would be it.
—Ron Hranac, Technical Leader, Cisco Systems and Senior Technology Editor, Communications Technology magazine

The authors have done a remarkable job of preserving the topical organization of the first edition, while introducing rapidly emerging technologies essential to the triple play of services. Following the methodology of its predecessor, the second edition provides enough information on topics such as VoIP and home networks to become conversant with the technology, and a significant set of references for those that must delve into the details of implementation.
—Justin J. Junkus, President, KnowledgeLink, Inc., Telephony Editor, Communications Technology magazine

This 'post-Modern' 2nd edition of their classic text is the new reference standard for the Cable industry.
—Dan Pike, CTO, GCI Cable

I have been in the Cable business since 1965 and over the years have read most of the "books" written about the technical aspects of cable systems. This book is the most comprehensive I have ever read. If you want to be successful as a "technologist" in Cable Communications today, this is a "must have" book.
—Dave Pangrac, President, Pangrac & Associates Consultants Inc.

Faced with constant changes in cable-broadband technology, our engineering community needs an engineering guide dealing with issues of today and the future, based on solid fundamentals that brought us into the 21st Century. This book is IT!
—Rex Porter, Charter/Senior Member, SCTE, and Vice-Chairman, Cable TV Pioneers

During my 30 years in the cable television industry, I have not encountered more qualified engineers than Ciciora, Farmer, and Large. For this second edition, digital expert Michael Adams joins the team and helps to create expanded coverage of advanced technologies. Now, this wealth of technical knowledge is available to everyone.
—William W. Riker Senior Vice President, Operations and Chief Technology Officer, The Cable Center

The Morgan Kaufmann Series in Networking

Series Editor, David Clark, M.I.T.

Modern Cable Television Technology: Video, Voice, and Data Communications, 2e
Walter Ciciora, James Farmer, David Large, and Michael Adams

Bluetooth Application Programming with the Java APIs
C Bala Kumar, Paul J. Kline, and Timothy J. Thompson

Policy-Based Network Management: Solutions for the Next Generation
John Strassner

Computer Networks: A Systems Approach, 3e
Larry L. Peterson and Bruce S. Davie

Network Architecture, Analysis, and Design, 2e
James D. McCabe

MPLS Network Management: MIBs, Tools, and Techniques
Thomas D. Nadeau

Developing IP-Based Services: Solutions for Service Providers and Vendors
Monique Morrow and Kateel Vijayananda

Telecommunications Law in the Internet Age
Sharon K. Black

Optical Networks: A Practical Perspective, 2e
Rajiv Ramaswami and Kumar N. Sivarajan

Internet QoS: Architectures and Mechanisms
Zheng Wang

TCP/IP Sockets in Java: Practical Guide for Programmers
Michael J. Donahoo and Kenneth L. Calvert

TCP/IP Sockets in C: Practical Guide for Programmers
Kenneth L. Calvert and Michael J. Donahoo

Multicast Communication: Protocols, Programming, and Applications
Ralph Wittmann and Martina Zitterbart

MPLS: Technology and Applications
Bruce Davie and Yakov Rekhter

High-Performance Communication Networks, 2e
Jean Walrand and Pravin Varaiya

Internetworking Multimedia
Jon Crowcroft, Mark Handley, and Ian Wakeman

Understanding Networked Applications: A First Course
David G. Messerschmitt

Integrated Management of Networked Systems: Concepts, Architectures, and their Operational Application
Heinz-Gerd Hegering, Sebastian Abeck, and Bernhard Neumair

Virtual Private Networks: Making the Right Connection
Dennis Fowler

Networked Applications: A Guide to the New Computing Infrastructure
David G. Messerschmitt

Wide Area Network Design: Concepts and Tools for Optimization
Robert S. Cahn

For further information on these books and for a list of forthcoming titles, please visit our website at http://www.mkp.com

Modern Cable Television Technology

Video, Voice, and Data Communications

Second Edition

WALTER CICIORA
Consultant

JAMES FARMER
Wave7 Optics

DAVID LARGE
Altrio Communications

MICHAEL ADAMS
Broadband Semantics



ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Morgan Kaufmann is an imprint of Elsevier

MORGAN KAUFMANN PUBLISHERS

Senior Acquisitions Editor Rick Adams
Associate Editor Karyn Johnson
Senior Project Manager Angela Dooley
Cover Design Ross Carron Design
Cover Photographs Fred Spencer and Carolyn Terry
Composition and Illustrations Integra Software Services Pvt. Ltd, Pondicherry, India
Copyeditor Elliot Simon
Proofreader Fran Pulver
Indexer Northwind Editorial
Interior printer The Maple-Vail Book Manufacturing Group
Cover printer Phoenix Color

Morgan Kaufmann Publishers in an Imprint of Elsevier
500 Sansome Street, Suite 400, San Francisco, CA 94111

This book is printed on acid-free paper.

Copyright © 2004, 1999, Elsevier Inc. All rights reserved.

Designations used by companies to distinguish their products are often claimed as trademarks or registered trademarks. In all instances in which Morgan Kaufmann Publishers is aware of a claim, the product names appear in initial capital or all capital letters. Readers, however, should contact the appropriate companies for more complete information regarding trademarks and registration.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopying, or otherwise—without written permission of the publishers.

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone: (+44) 1865 843830, fax: (+44) 1865 853333, e-mail: permissions@elsevier.com.uk. You may also complete your request on-line via the Elsevier homepage (http://elsevier.com), by selecting "Customer Support" and then "Obtaining Permissions."

Library of Congress Catalog Number: 2003107478
ISBN: 1-55860-828-1

For all information on all Morgan Kaufmann publications
visit our website at *www.mkp.com*

Printed in the United States of America

03 04 05 06 07 5 4 3 2 1

To Our Families

Foreword

Archer S. Taylor *(retired)*
Co-founder, Malarkey-Taylor Associates

The first edition of *Modern Cable Television Technology* was so well received by engineers and technicians that additional printings were needed to meet the demand. Since the publication of the first edition, advances in the application of modern information technologies have transformed the Cable Television industry at an astonishing rate. This second edition has been substantially revised to enable the reader to catch up with recent advances that have become commonplace in modern cable networks, and to prepare for new developments to come.

Virtually every household in the United States, and many businesses, already have access to multi-channel television service distributed over Cable networks. Comparable facilities are also in place throughout the world, although not generally as universally accessible as in the U.S. Multi-channel television still retains its traditional role as the "cash cow" supporting the ubiquitous coaxial cable and optical fiber networks. Yet, in the public view, high speed broadband Internet communication seems on course to surpass entertainment as the dominant function of modern cable networks. Moreover, the cable industry is becoming a significant participant in the traditional switched telephone business, adding voice to video and data communication services. Even the trade associations now describe the expanding scope of the industry as "Cable Telecommunications."

The first edition provided the most comprehensive and authoritative presentation yet available regarding the physics and practical engineering aspects of coaxial cable and optical fiber networks. This work is still valid, representing the essential foundation for an engineer or practitioner in the field today. With the second edition of *Modern Cable Television Technology,* the authors expand on the authentically expert treatment of these networks with in-depth coverage of the technologies that are currently transforming the medium from its original function of distributing television signals into a major telecommunication medium serving the broader requirements of data transportation.

An excellent introduction to advanced technologies was included in the first edition. In the second edition, the authors have substantially revised and expanded their coverage of developments that have migrated out of the laboratory

into everyday operations only within the past few years. For example, high speed cable modems are now being connected to the Internet at an accelerating pace. Digital technology is rapidly overtaking analog for television distribution as well as voice and data communications, and digital compression has revolutionized bandwidth efficiency for non-video as well as television. Optical fiber lines, formerly used mainly for hub interconnections, now comprise the primary transport medium for analog video as well as digitally modulated signals, while coaxial cable serves the "last mile" into subscriber premises. Wavelength division multiplexing, no longer a laboratory experiment, is now deployed as a practical operational component of optical network architecture. Internet protocol (IP) telephony is on the verge of operational deployment, and home networks have become essential features in new construction. Finally, technical standards are being generated and revised with amazing frequency to ensure that the remarkable functions of the modern cable network are able to operate together without conflict.

For the second edition, Michael Adams joins Walter Ciciora, James Farmer and David Large as co-author, with important contributions. New sections on the physics and practical engineering aspects of digitally compressed television and home networking reflect his special expertise in telecommunications development, stemming from nearly a quarter century's work with Time Warner, Nortel, and Digital Equipment (DEC). The entire outstanding team of authors brings to the preparation of this second edition a wealth of experience based on productive professional careers working with the development, deployment and operation of recent developments in cable communications.

The authors are to be commended for so effectively updating this important engineering reference work. The second edition of *Modern Cable Television Technology* represents a very substantial revision of the first edition, including virtually all of the material from the original edition. For engineers and technicians engaged in the modern cable telecommunication environment, the second edition is the essential guide to the latest developments and preparation for what is to come.

Preface

Overview

When we completed the first edition of *Modern Cable Television Technology*, we described the major characteristic of the technologies covered as *accelerated change*. That has proven to be true, with so many major changes taking place since 1999 that a second edition was required to catch up. Even the book's title is really obsolete, since so-called "cable television" systems have moved far beyond simply transporting video entertainment to the development of scalable-bandwidth, multi-service platforms supporting virtually every wired communications service desired by individuals and small businesses.

Among other changes since the first edition was published have been great progress in standards and practices for delivery of high speed data, home networking, IP-based voice, optical dense wavelength division multiplexing, new video compression techniques and integrated voice/video/data transport. With the help of a fourth author and a lot of industry professionals, we have attempted to update the previous material and add new material on emerging technologies on the same consistent, applications-oriented level as the first edition. Our aim has been to give the working engineer the information needed to understand the fundamentals of all the technologies used to build and operate a cable television system, while avoiding the detailed minutiae best left to designers of the hardware and software involved.

Intended Audience

The target audience includes several constituencies. Most important are our colleagues in the industry, particularly the members of the Society of Cable Telecommunications Engineers (SCTE). It is our intention that this book serve not only as a day-to-day reference, but also as the most useful tool for those preparing for the SCTE's certification examinations. The cable industry is made up of those who build cable systems, operate them, connect to them, and supply the equipment used in them. All are part of the intended audience. The consumer electronics industry produces products that are connected to these systems. A better understanding of cable technology can facilitate equipment and system designs that will interwork seamlessly to the mutual benefit of operators, manufacturers and end

users. An ever increasing percentage of the information that moves through cable systems is digital and cable system staffs now typically include technologists whose previous experience is limited to data networks. We hope to provide a mutual education vehicle between those whose backgrounds are in the broadband, linear world and those who think of information in a baseband, digital format. Similarly, the traditional telecommunications carriers heavily use optical fiber to carry digitally modulated signals. We hope to provide a tool for telecom engineers to understand the critical differences when linearly modulated signals are carried over optical networks. Others who should find material of relevance include operators of campus networks, regulators seeking a better understanding of the industry, telecom operators seeking to expand their service offerings, the emerging fiber-closer-to-the-home network architects, and private network operators. Lastly, we hope that the general technical fraternity might find satisfaction for a curiosity about the technology behind cable television.

Although we have not dealt thoroughly with metric measurements, PAL and SECAM analog video formats, and non-U.S. regulations, the technology differences are relatively minor, and most of the material should be of interest and value to non-North American audiences.

Fitting the material to the audience is a challenge when readers come with widely different experiences and pre-existing knowledge bases. Some will be experienced engineers from related fields who are unfamiliar with the details of cable television transmission technology, but intimately familiar with television. Others will be familiar with various methods of transporting signals, but not with some of the signal types carried over our networks. Still others will be experienced with radio frequency techniques, but not with optical transport. Finally, some will be equipment designers who are intimately familiar with their market niche, but need an understanding of how and whether their products fit into the business of cable television. Though some will be engineers with graduate degrees, others will be self-taught. Even those who have previously studied these topics may benefit from the refreshing.

Another aspect of the problem is to fit the scope of interest. Some may find the history of the industry fascinating, while others have no interest at all. Some want to know details, others want the shortest path to useful results. Our task was to organize and present the material at a level that is most useful to the majority of readers, while providing references to more-detailed treatments for those needing it. We also faced an ultimate restraint on the size of the book — you have to be able to lift it without requiring assistance! Our advice is to boldly skip over what is beyond your interest, use referenced material for more details where needed, and seek us out at the next convention to ask about what is missing.

Approach

The general level of the book is intended to be appropriate for a competent technologist who understands the basics and wants to understand cable tele-

vision technology or who already understands some aspects of the business, but needs to expand his scope. Mathematics is held to the minimum needed to understand apply the technologies presented, rather than to rigorously prove each theory, with the hope of building a practical and useful understanding of the technology. It is not our purpose to develop concepts from first principles or to be scientifically "pure." In that regard, we have tended to use common engineering units, as opposed to scientific units, in formulas even when that results in the inevitable correction factors. Our intention is to create a text that will have lasting value in an era of rapid technological change. This requires an emphasis on the fundamentals that are the basis upon which the advanced technology is built.

The general flow of the material here roughly parallels a "downstream" flow towards the end user. We begin with the signals, then consider the headend where those signals are received or generated and processed, then the transmission system that carries them throughout the served community. Finally, we deal with the terminal equipment and various compatibility and security issues involved.

Content

Cable television networks were originally designed for the purpose of delivering analog, broadcast-format television to widely dispersed viewers with a quality level that competed with alternate technologies and met FCC requirements. As other signals have been added to these networks, they have benefited by being transmitted through the very clean, low-distortion channels required for analog television, but also been constrained by the historically-based requirement of conforming to the 6-MHz television channelization. Thus, the book begins with a brief history of the industry, followed by analog television. It then deals with other signal formats and modulation types. With that background, it then deals with the central signal processing center (the "headend"), the distribution systems, and the subscriber end of the network.

The book is organized into six major parts:

Part 1: Once Over Lightly

Part 2: The Signals

Part 3: Headends

Part 4: Broadband Distribution Systems

Part 5: System Architecture

Part 6: Customer Interface Issues

The parts, and even the chapters, stand alone reasonably well. Dependencies of material from other chapters are clearly indicated. The reader having specific needs or a generalized knowledge of cable is invited to dive directly into the

chapters of interest. A newcomer to the industry is advised to approach the process iteratively. First a light treatment, then more detail from a sequential reading of the chapters. Even so, skipping a chapter of little interest shouldn't create a serious continuity problem.

Part 1, *Once over Lightly*, consists of a single chapter that gives a casual review of cable technology, business, regulation and history. This should be easily readable, even for those with no technology background.

Part 2, *The Signals*, covers in five chapters the technical details of the signals currently being transmitted over cable networks. These networks are inherently service-agnostic, however, so a reading of this section will also serve to inform the creator of a new service class about the transmission channel parameters to be expected. *Chapter 2, Analog Television*, covers the signals that drove the early development of the industry (and the only signals whose minimum performance level is specified by the Federal Communications Commission). The technical history of analog television is traced from the original black-and-white format through the additions of color, stereo sound and ancillary signals. *Chapter 3, Digitally Compressed Television*, covers the processes of digitizing the original analog signal, compressing it by removing redundancies, formatting it, and preparing it for transmission. Both standard-resolution and high-definition formats are covered. *Chapter 4, Digital Modulation*, covers all aspects of modulating an RF carrier with a baseband digital stream so that it can be transmitted through a shared, broadband, linear distribution network. This is an important chapter for those with a baseband, digital background, as the transmission channel conditions are considerably different in cable's frequency-division-multiplexed (FDM) environment when compared with the traditional telecom time-division-multiplexed (TDM) channels. *Chapter 5, Cable Networking Protocols*, covers all aspects of transmitting high-speed data through cable networks. In particular, the Data Over Cable System Interface Specification (DOCSIS) is covered in detail, as are non-DOCSIS protocols of interest to cable operators. *Chapter 6, Cable Telephony*, covers the technology required to carry digitized voice signals over these shared networks. Both proprietary switched-circuit TDM and newer Voice-over-IP (VoIP) protocols are covered.

Part 3, *Headends* covers, at a system engineer level, the originating and processing of signals in the headend. The three chapters cover reception, processing and headend operations. *Chapter 7, Signal Reception*, covers all aspects of receiving signals, including satellite, off-air and direct cable or fiber paths. *Chapter 8, Headend Signal Processing*, covers analog and digital modulation, handling of off-air digital signals (which are in a different modulation format than usually used for cable digital signals), ad insertion and server-based services such as video-on-demand (VOD). Chapter 9, Headend Operation, covers the overall RF spectral planning, RF and Optical signal management in the headend, and signal quality tests.

Part 4, Broadband Distribution Systems, covers in seven chapters the technologies used to transport signals throughout service areas. *Chapter 10, Coaxial RF Technology*, covers the linear, RF technology that forms the "last mile" of the hybrid fiber/coaxial (HFC) networks that are almost universally the backbone of modern cable systems. *Chapter 11, Coaxial Distribution System Design*, shows how that technology is used to design and specify coaxial networks. *Chapter 12, Linear Fiber-Optic Signal Transportation*, discusses the basics of transporting RF FDM spectra through use of linearly modulated optical transmitters. *Chapter 13, Wavelength Division Multiplexing*, expands on Chapter 12 to discuss and calculate the tradeoffs when transmitting multiple optical signals on different wavelengths of light through shared fibers. *Chapter 14, Broadband Amplitude Modulated Microwave Links (AMLs)*, covers the use of microwave to transmit wideband FDM spectra where the use of cabled media are not practical. *Chapter 15, End-to-End Performance*, covers the total cascaded performance of the optical-plus-RF distribution path, with an emphasis on downstream transmission. *Chapter 16, Upstream Issues*, covers the unique aspects of signals that are transmitted from end users towards the headend.

Part 5, *System Architecture*, covers in four chapters how the network elements go together to form a complete system, starting with requirements, through design principles to expected reliability. *Chapter 17, Service-Related Architecture Requirements*, examines each current or contemplated cable-delivered service and examines its requirements in terms of signal quality, bandwidth, scalability and reliability. *Chapter 18, HFC Architectural Elements and Examples*, covers the principles of service-related overall design, including how the headend and distribution system interrelate. It gives examples of various architectural alternatives. *Chapter 19, Fiber-Deep Systems*, covers emerging network architectural alternatives in which all, or nearly all, the distance from headend to subscriber is through optical fibers. *Chapter 20, Network Reliability and Availability*, covers the process of designing distribution systems to meet defined goals for failure rate and availability.

Part 6, *Customer Interface Issues*, covers in five chapters issues associated with delivering services to customers at their homes. *Chapter 21, Program Denial Technologies*, covers the processes used to selectively deliver analog video programming to subscribing customers, while protecting them from unauthorized reception. *Chapter 22, Digital Video Reception*, covers the issues related to real-time customer interaction with headend equipment that delivers digital video programming. *Chapter 23, Consumer Electronics Interface*, covers the operational issues that arise when customer-owned equipment is connected to cable systems, while *Chapter 24, Equipment Compatibility*, covers the technical compatibility issues. *Chapter 25, Home Networks*, covers issues of interconnecting data equipment within customers' homes.

In addition to the text, various appendices are included for reference. *Appendix A, Channel Allocation*, is a partial reprint of CEA/ANSI 542-B, the

U.S. standard for downstream channelization in cable systems. *Appendix B, Video Waveforms* is based on EIA/TIA 250C and contains details of analog NTSC video waveforms and timing information. The *Glossary* contains an alphabetized listing of acronyms used throughout the book and should remove the necessity for a reader to search through the text for the first usage of a term when coming across it later in the book.

Background

We, the authors of this book, have written hundreds of technical papers and trade magazine articles, delivered countless presentations, and presided over dozens of technical conference sessions during combined activities exceeding a century in the cable industry. Some of us have also taught and lectured both within and outside the industry. In the process, we have certainly developed individual styles. We also have overlapping, but distinct, areas of expertise. Those who know us well can thus make a good guess as to which author made the primary contribution to each chapter. Every chapter, however, has had the benefit of mutual author review, independent review by outside authorities in each subject matter, and stylistic review by Morgan Kaufmann's own copy editors. Hopefully, the result has a consistent feel for the reader.

The seed of the first edition was a monograph written by Walt Ciciora when he was with American Television Communications (the predecessor of Time Warner Cable). After extensive revision, that became a CableLabs publication. Morgan Kaufmann Publishers asked if it could be expanded into a book. Walt initially enlisted Dave Large to assist, and when the effort was about to overwhelm the two of them, they shanghaied Jim Farmer and among them, the First Edition was born with little resemblance to its ATC/CableLabs parent. Published in late 1999, it has been very well received by the cable technical community and was honored as the recipient of the Cable Center's 2000 Cable Book Award.

Things are changing quickly in cable-related technology, however, and by mid-2001 we were already discussing the need for a second edition and sketching a revised table of contents. This time we were happy to add Michael Adams to the team (especially because of his expertise in digital video), while Walt Ciciora has served as a general consultant, rather than an active author. Chief among new material that needed to be covered were the emerging digital standards for voice, data, video and multi-media. On the distribution side, wavelength-division multiplexing had become a mainstream technology and several vendors were proposing various fiber-deep architectures. At the same time, a few existing technologies were beginning to fade, among them analog scrambling methods. We hope you find that we have achieved a balance between covering everything that needed to be included in a comprehensive book on cable technology, while turning out a book that is still lighter than the average spectrum analyzer!

Acknowledgments

Given the breadth of cable technology, there are no four people who can hope to be sufficiently familiar with it all to have written this book. We have relied on both the published materials of many others and, on a more direct level, the second edition has benefited, as did the first edition, from the specific contributions and review of many. We have, we hope, properly credited information gleaned from previously published works.

Special acknowledgment goes to Mary Philips for advice, draft review and use of her published material, all of which greatly aided our treatment of dense wavelength-division multiplexing. New material was contributed and reviewed by Erich Arnold (Arris Interactive), Tom Carr, (Wave7 Optics), Bob Gaydos (Concurrent Computer), Dean Stoneback (Motorola), and Ned Mountain and Louie Lividitis (both Wegener Communications). Ned even picked up the check for lunch! John Kenny (now Wave7 Optics), who was most helpful in the first edition, again contributed his proofing ability (and his insistence that power, not energy, is present at an interface). Kinney Bacon and Tony Wasilewski (both Scientific-Atlanta), Julius Bagley (consultant), and Jeff Reidmiller, (Dolby Labs) contributed their expertise. Consultant Charles Poynton was preparing his Morgan Kaufmann book about digital video while we were writing. He excerpted it for a tutorial at the IEEE International Conference on Consumer Electronics, which was so good he found himself pressed into service answering a lot of questions for us, and then proofing one chapter.

The cover art for the second edition originated with a series of pictures taken by Fred Spencer and Carolyn Terry of Altrio Communications, who are not only superb technologists, but competent photographers.

Much of the material from the first edition, while updated, has been retained largely intact. That edition reflected the contributions of many, including the following proposal reviewers (with their then-current affiliations): Alex Best (Cox Cable), Rex Bullinger (Hewlett Packard), Keith Cambron (Pacific Bell), Jim Chiddix (Time Warner Cable), Brian Colbry (Bell Labs), Joseph Glaab, (General Instrument Corporation), Andrew Glasspool (Farncombe Technologies, UK), Paul Shumate (Bellcore), and Archer Taylor (The Strategis Group).

Chapter reviewers included: George Abe (University of California, Los Angeles), Tim Brophy (General Instruments Corporation), Bob Dickinson (Dovetail Surveys, Inc.), John Carlucci (MystroTV), John Coppola (Cox Communications), Craig Cuttner (HBO), Keith Fleming (Antec), Joseph Glaab, Kenneth Gould (Time Warner Cable), David Grubb (General Instrument Corporation), Dan Harris, PhD (Corning Glass), Ted Hartson (Scottsdale Television Labs), Jud Hoffman (Panasonic), Steve Hughey (Antec), Doug Jones (U.S. West), John Kenny (Antec), Keith Kreager (Antec), Bill Lagoni (Thomson Consumer Electronics), Tom Lookabaugh (University of Colorado at Boulder), Michael Maslaney (Arris Interactive), M.C. "Mitch" Matteau (Arris Interactive), Marv Nelson (SCTE), Jim Raiser (Scientific Atlanta), Gerald Robinson (Scientific Atlanta),

Blair Schodowski (Antec), Oleh Sniezko (TCI Cable), S.V. Vasudevan (BigBand Networks), Merrill Weiss (Merrill Weiss Group), and Norman Woods (Cable AML Inc.).

Special acknowledgments went to: Andrew Large, for the treatment of user-perceived reliability; Nick Hamilton-Piercy, for the contribution of recent failure analysis data; Ted Hartson, for additional help on the fine points of the history of television technology; and Dr. Francisco Bernues of Cable AML for providing the approach used for estimating microwave rain fade outage probability. For help with various technical details, we were grateful for the assistance of: Vic Williams and Dave Womble (Scientific Atlanta), Scott Ramsdell and Keith Kreager (Antec), John Vartanian (Viewers' Choice), Mitch Matteau and Mark Bugajski (Arris Interative). Archer Taylor was a remarkable contributor, both as a reviewer, and in sharing his extensive knowledge and experience in the industry. He provided a number of interesting historic notes and also suggested a notation for the addition of cascaded noise and distortion in distribution systems.

The first edition was shepherded through Morgan Kaufmann by Jennifer Mann who was editor, friend and, occasionally, stern taskmaster, aided by Karyn Johnson our editorial assistant and Cheri Palmer, our production editor.

Rick Adams of Morgan Kaufmann spearheaded the effort that led to the second edition, working with the authors to architect a revised book that covered the new material, while staying within a reasonable size constraint. Karyn Johnson has done the heavy lifting of bringing together all of the updated and new material and overseeing the process of making a cohesive whole out of it. Angela Dooley and Elliot Simon had the unenviable task of fixing all the authors' grammatical and style errors. As always, the folks at MKP have been an author's dream to work with.

Our heartfelt thanks to all the contributors to both editions. If we have failed to thank someone for his or her assistance (or misspelled a name or publication, despite our best efforts), we apologize. If we have misinterpreted your suggestions or stubbornly refused to take your good advice, we take full responsibility for the result.

Lastly, we wish to thank those in the industry we have been privileged to know, our colleagues, and those who have gone before us and left us with a rich legacy of tools with which to further build. We have enjoyed the adventure and look forward to years of further challenges.

Acknowledgments

Chapter 1 figures are based on Walt Ciciora, *Cable Television in the United States*. May 1995. Cable Television Laboratories, Inc., Louisville, CO. © 1995 Cable Television Laboratories, Inc.

Tables 3.2 and 3.3 are based on ANSI/SCTE 43 2001 (formerly DVS 258) *Digital Video Systems Characteristics Standard for Cable Television*, pages 2 and 3, 2001. Society of Cable Telecommunications Engineers, Exton, PA. © Society of Cable Telecommunications Engineers, Inc.

Figures 3.7(a), 3.7(b), 3.8, 3.10, 3.11, Table 3.4, and Figure 3.19 are based on S. Merrill Weiss, *Issues in Advanced Television Technology*. Pages 33, 73, 74, 90, and 141, 1996. London: Focal Press. © 1996 Focal Press.

Figures 3.7(c), 3.5, and 3.9(a) are based on Keith Jack, *Video Demystified*. Pages 517, 518, and 519, 1996. Eagle Rock, VA: HighText Publications. © 1996 LLH Technology Publishing.

Figure 3.17 was provided by Neil Brydon of Harmonic Inc.

Figures 3.16, 3.20, and 3.21 are based on *Guide to the Use of the ATSC Digital Television Standard*, Doc. A/54. Pages 24, 68, 86, 97, and 146, October 1995. Advanced Television Systems Committee, Washington, D.C. © 1995 ATSC Organization.

Figures 3.22, 3.24, and 3.26 are from Michael Adams, *OpenCable Architecture*. Pages 64, 65, and 66, 2000. Indianapolis: Cisco Press. © 2000 Cisco Press.

Figures 3.23 and 3.25 are based on SCTE 40 2001 (formerly DVS 313), *Cable Network Interface Specifications*. Pages 12 and 16, 2001. Society of Cable Telecommunications Engineers, Exton, PA © Society of Cable Telecommunications Engineers, Inc.

Figure 3.27 is based on ANSI/SCTE 67 2002 (formerly DVS 379), *Applications Guidelines for SCTE 35*. Page 5, 2001. Society of Cable Telecommunications Engineers, Exton, PA. © Society of Cable Telecommunications Engineers, Inc.

Figure 3.29 is based on John Watkinson, *An Introduction to Digital Video*. Page 218, 1994. London: Focal Press. © 1994 Focal Press.

Figure 10.12 is based on data furnished by Howard Carnes of Antec Corporation.

Figure 12.6 is based on Ronald C. Cotton, *Lightwave Transmission Applications*. Page 108, September 15, 1993. The graph is used with permission of CommScope, Inc., Cable Television Laboratories Inc., Louisville, CO. © 1993 Cable Television Laboratories, Inc.

Figure 12.7 is based on data furnished by CommScope, Inc.

Figures 12.10 and 12.11 are based on Dogan A. Atlas, Fiber Induced Distortion and Phase Noise to Intensity Noise Conversion in Externally Modulated CATV Systems, 1996 *National Cable Television Association Technical Papers*. Page 291 and 292, April, 1996. National Cable Television Association, Washington, D.C. © 1996 National Cable Television Association.

Figure 12.13 is based on Dan Harris, Primer on Baseband Digital Transmission on Single-Mode. Page 94, August 1996. *Communications Technology*, Potomac, MD. © 1996 Phillips Publishing International, Inc.

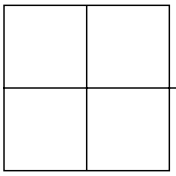
Tables 14.1 and 14.2 are adopted from material furnished by Cable AML, Inc., September 5, 1997. Cable AML, Inc., Torrance, CA.

Table 14.3 and the supporting text and equations are based on material furnished by Dr. Francisco Bernues of CableAML, Inc. September 5, 1997. CableAML, Inc., Torrance, CA.

Figures 14.4, 14.5 and 14.6 are reprinted with permission from the International Telecommunication Union. ITU-R Recommendation ITU-R PN.837-1, *Characteristics of Precipitation for Propagation Modeling*. Figures 1, 2 and 3. © 1996 International Telecommunication Union. Telecommunications Union, General Secretariat, Sales and Marketing Service, Place des Nations, CH-1211, Geneva 20, Switzerland.

Appendix A is based on EIA/ANSI-542, *Cable Television Channel Identification Plan*. Electrical Industries Association. Available from Global Engineering Documents, 800-854-7179. Used with permission of Global Engineering Documents.

Appendix B is based on EIA/TIA-250C, *Electrical Performance for Television Transmission Systems*. Electrical Industries Association. Available from Global Engineering Documents, 800-854-7179. Used with permission of Global Engineering Documents. Also based on Margaret Craig, *Television Measurements: NTSC Systems*. Pages 76–78. Tektronix, Beaverton, OR.



Chapter 1

Introduction to Cable Television

1.1 Introduction

Cable television is an industry and a technology that has outgrown its historical name. Modern "cable television" networks are used to provide a wide range of services, including analog and digital video, digital audio, high-speed data, and telephony.

The essential distinguishing characteristics of cable television networks are that they include broadband (typically 0.5–1 GHz of total bandwidth), highly linear distribution systems designed to carry many modulated radio frequency (RF) signals with a minimal amount of mutual interference between a central point and many customers, where signals are delivered via coaxial cables to and from terminal equipment. Because of these characteristics, the networks are service-agnostic to the extent that they will carry any information that can be modulated on a compatible RF carrier. Modern cable television networks are almost always two-way, use optical fiber extensively, and are segmentable so as to allow simultaneous frequency reuse in various network sections.

Historically, the cable television business was based exclusively on delivery of television programming, and it has been very successful in that regard. As of 1999, nearly 97% of U.S. television households had cable television service available, and approximately 66 million households subscribed to at least the lowest tier of video service, representing almost 67% of U.S. television households.[1] Those levels have changed only slowly over the past few years.

Because cable television has been so successful and has enjoyed such vigorous growth and acceptance, it has spawned video competitors, including prerecorded media, direct broadcast satellite (DBS), video streaming over the Internet, as well as the interest of the telephone industry. Increasing revenue streams from connected households, made possible by multiple service offerings, has also changed the economics of the industry sufficiently that customers in some markets now have a choice between two cable television operators who have constructed parallel distribution networks serving the same homes. In the common terminology of the cable industry, the company building the second network is known as an *overbuilder*. Overbuilders may offer service as a second, franchised cable television operator or as an *open video system* (OVS) operator.

Taking advantage of new technologies and, in particular, the falling prices of electro-optical components, these networks typically use optical fibers to carry signals closer to subscribers than legacy cable operators, in some cases all the way to subscriber homes.

In the world of high-speed data communications, and Internet access in particular, services offered by cable operators have codeveloped with other wired and wireless options. For residential users, Internet access was historically provided almost exclusively through dial-up modems directly to *Internet service providers* (ISPs). Many applications, however, run discouragingly slowly at the data rates that are possible through standard telephone connections, and this has led cable companies to develop connection services that are 10–100 times faster. In the competition among broadband service providers, cable has outsold its competitors by about 2:1, with a telephone technology known as *digital subscriber line* (DSL) providing the strongest competition to date. Competing satellite and wireless terrestrial data transport technologies are still developing market share.

In offering voice telephone service, it is the cable operator who is the overbuilder, since residential telephone service was available to virtually every household in the United States before cable television operators entered that market segment. When offered by cable television operators, telephone service is regulated by the same agencies who regulate the incumbent telephone companies, leading to completely different regulatory authorities for services that share the same physical network.

Cable-offered telephony comes in at least two technical versions and two product classifications. Initially, telephone offerings utilized dedicated signal-processing equipment that was carefully engineered to meet the high reliability expectations of this service. As of early 2003, that type of equipment was still used to service the large majority of installed telephone customers. The newest version, known as *voice-over-IP* (VoIP), shares terminal equipment with the high-speed data service, offering the potential for reduced equipment costs.

Regardless of how the signals are handled technically, cable operators may offer a *primary-line* service or only a *secondary-line* service. Primary-line service competes directly with the incumbent telephone operator for all residential telephone business, but it requires that the cable network and equipment meet the reliability standards to be the "life-line" communications link for customers in the case of an emergency (while not mandated by regulatory agencies, this voluntary requirement is generally defined as a service that is available at least 99.99% of the time). Secondary-line service cedes the first line in each home to the incumbent but competes for additional lines. The assumption is that the availability needn't be as high, saving the cable operator capital upgrade cost and allowing it to offer a lower price.

In the following chapters, you will gain a solid understanding of the technologies required to deliver broadband services to and from homes. You will see how the pieces fit together to make up a complete system for the transmission of

information and entertainment choices to consumers. If yours is a related business, you will better understand how it fits with the cable industry. If you are already knowledgeable in some aspects of the broadband networks used by cable, this book will fill in the gaps.

1.2 Technology-Related Regulatory Issues

Cable's high visibility has attracted the attention of regulators and legislators. Important public policy issues are involved. Understanding what cable television is, how it works, and its economics will help decision makers in these areas. This understanding will also aid network designers in determining which technologies are appropriate for cable television applications and which are not. One aim of this book is to aid in that understanding.

The issues are fundamentally different from those that govern incumbent telephone companies, broadcasters, or DBS operators because of the nature of the cable television business model. Unlike telephone companies, which are fundamentally providers of bandwidth but have no interest in content, and DBS operators, which are fundamentally purveyors of content and for which the network is just a private delivery mechanism, cable companies are involved in construction, network operations, bandwidth leasing, passive retransmission of certain signals, resale of certain programs and signals, equipment leasing, and program generation. These sub-businesses are the result of cable's need to define itself over the years in a very entrepreneurial way and make it hard to pigeonhole it into a defined niche. Following are a few of the regulatory issues that are affected by the technology of cable systems.

1.2.1 Access to Networks by Unaffiliated Third Parties

Current regulatory policy allows unaffiliated competitive local exchange carriers (CLECs) to offer telephone and data services over physical plant constructed by incumbent telephone companies and to access incumbent telephone company facilities for the purpose of installing the necessary equipment. In some cases, signals from the two companies may share the same physical wires between a telephone switching office and customers' homes. A recurring regulatory question is whether, and to what extent, access to cable systems should be granted to such third parties. Should access be granted at the physical level (shared wires), at the bandwidth level [allocated radio frequency (RF) channels in the network], or at the service level (where the third party's signals are processed by equipment owned and operated by the incumbent cable operator)?

The question of third-party access potentially arises in connection with all types of cable television services, often as a result of the vertical organization of cable companies, which not only build and operate networks but may also invest in companies providing video programming, operate their own ISPs, resell

long-distance telephone service, lease various types of terminal equipment, and have other relationships that may cause them to favor some suppliers over others.

The technical issues are much more complex in the cable television world than in the telephone world because cable systems are not passive, but, rather, include electro-optical and RF active equipment whose operation requires very tight control over signal parameters. Furthermore, cable system operators are responsible for inadvertent radiation of their signals, which can be caused by poor workmanship on the part of a third-party installer.

1.2.2 Cable Operators' Responsibilities in Transition to Digital Broadcast Television

The eventual transition of over-air broadcast television from analog to digital (including both standard and high-definition formats) is a certainty, though the timing is not. The practical issues are many.

- Broadcasters are reluctant to produce high-definition television (HDTV) in quantity because of the high costs involved for a currently-very-limited viewing audience, yet HDTV programming in sufficient quantity will be required to encourage consumers to purchase digital receivers.
- The costs of digital receivers remain high due to limited demand as well as inherently higher production costs.
- DBS and cable television operators offer both standard-definition television (SDTV) and a limited quantity of HDTV programming, but using formats that are incompatible with each other and with over-air broadcast digital. Even among cable network operators and between satellite operators, receivers are not currently interchangeable. Furthermore, HDTV programs use four to six times the bandwidth of SDTV programs, so any further migration to HDTV on the part of DBS and cable operators will inevitably be at the expense of reduced program variety, given a finite amount of available bandwidth.
- Given that most households in the United States receive their television programming from a cable television system, broadcasters maintain that a successful transition is dependent on cable operators' carrying their analog and all their digital signals during the transition period. Given their limited bandwidth, the relative scarcity of deployed HDTV receivers, and the duplication of much of that programming, cable operators believe that this is an unreasonable burden on their limited bandwidth and an imposition on the editorial control of their systems; they point out that the immediate result will be to reduce programming choice for their customers and to deny access to some nonbroadcast programmers.
- For the consumer, the result has been that those wishing access to digital programming typically purchase a display-only receiver, then mate it with separate satellite, cable, over-air, and/or physical media players, as required.

1.2.3 Market Power Issues

Although not a legal monopoly, cable television systems, in total, serve about two-thirds of U.S. households with at least television service. This has led some regulators to regard them as *de facto* monopoly providers and to enact regulations to limit their power to control various aspects of the marketplace.

Consumer Terminal Equipment

The Cable Act of 1996 led to Federal Communications Commission (FCC) regulations in 1998 that required that any required terminal equipment be accepted from unaffiliated third parties by July 2000. Note that this regulation applies only to cable television (video) services, not to telephone services, which are provided under separate regulatory framework. The status of data services is still unclear.

The technical issues in successfully moving terminal equipment into the retail market are many and include the following.

- The requirement that such devices operate under the secure control of the network operator in allowing selective access to programming present at its input terminals. In practical terms this requires that the decryption/descrambling system, or at least a portion of it, be supplied by and controlled by the network operator.
- In order to meet policy needs to support the transition of the over-air broadcast industry to digital, as well as to provide reception of both broadcast and narrowcast programming over the cable network, the terminals must be capable of receiving, demodulating, and decrypting multiple types of signals.
- In order to protect the cable network, any upstream signals (transmitted from the customer back to the operator for control of interactive services) generated by the receiver must be tightly controlled and conform to industry standards for interactive communications. As of mid-2003, no such complete set of standards existed.
- In order to protect the cable network from destructive interference and to protect over-air communications services from interference, such receivers must be adequately shielded to prevent ingress or egress of signals.
- In order to ensure operability in any cable system, the reception performance must be consistent with the signals delivered by advanced cable system architectures.
- In order to ensure adequate functionality, the receivers must include an adequate set of features to allow the required subscriber interactivity, such as purchasing on-demand programming or executing e-commerce transactions.

- Even if all these goals are met, the result will be that cable operators will be inhibited from offering new and innovative services if those services require an upgrade or enhancement to subscribers' terminal devices.

Although multiyear intensive negotiations between the cable television and consumer electronics industries produced the outlines of specifications for "digital-ready" television sets that could be sold through retail outlets and that would function as both an over-air and cable receiver,[2] the reluctance of the FCC to set and enforce adequate labeling, feature, and performance standards left the effort in limbo.

A more recent private agreement between seven major cable operators and 14 consumer electronics manufacturers covered only noninteractive receivers and, as of the writing of this book, had not yet been adopted by the FCC.[3]

The cable industry is rapidly moving all premium video services from older analog scrambling technology to all-digital transmission and encryption. Therefore, it is generally considered prudent to concentrate on the regulation of digital video and to let proprietary analog scrambling systems die a natural death.

Greater success has been reached in high-speed data service, where a durable industry-wide standard has been adopted for terminal equipment. The Data Over Cable Systems Interface Specification (DOCSIS) was originally a proprietary document, but it has since been recognized by the International Standards Organization (ISO). This evolving effort is in its third major release, and devices meeting DOCSIS are usable in the vast majority of cable systems in the Americas. A major contributor to the success of this effort has been the activities of the U.S. research and development consortium, Cable Television Laboratories (CableLabs), acting as the official certifying agency for new equipment. Perhaps the strongest indicator of the success of DOCSIS is that the price of data modems has dropped by a factor of four in about two years, while the price of proprietary cable television video terminal equipment has declined by less than 50%.

Access to Programming by Competing Network Operators

At the other end of the network, regulators are concerned that cable's market power may allow operators to effectively control access to video content that is, after all, the core product customers purchase. This control can take the form of exclusive deals covering certain geographic areas or unreasonably unequal rates. In many cases, large multisystem operators (MSOs) are vertically integrated to the extent that they hold major investments in programming suppliers and thus influence how and to whom the programs are made available.

In the interest of encouraging competition, current regulations forbid exclusive access to any programming that is distributed via satellite, allowing both DBS operators and overbuilders access to the vast majority of programming.

Service Bundling and Buy-Through

Cable operators often bundle certain services (that is, offer certain services only in combination with other services) or require customers to purchase certain services in order to have access to purchasing additional services. In some cases, this is required by regulations (such as the requirement that the lowest level of video service include all the local broadcast stations) and in some cases it reflects available technology (such as the inability to reasonably block the reception of some services while allowing access to others). In other cases, however, it reflects marketing decisions, contracts with certain program suppliers or strategic alliances that may not result in choices customers would have preferred. Regulators, therefore, need to know what is technically reasonable, as well as to assess what is a reasonable degree of constraint on cable operators' rights to control their product.

One example that reflects the development of the business is the bundling of ISP service with high-speed data transport service. In some cases, a high-speed ISP company actually operates the Internet access business for the cable operator, and the combined revenue for modem rental, ISP services, and transport are what supports the business. Allowing subscribers to choose a different ISP may be technically difficult, depending on the way the system is configured, but it may also cause a redefinition of a formerly integrated business segment.

1.2.4 Other Technical Regulatory Issues

Numerous other regulatory issues are related to technology, including the following.

Delivered Signal Quality

Current regulations cover only analog video signals and do not address any digital services or two-way communications. The analog regulations are intended to ensure a reasonable subjective quality of television viewing.

System Shielding Integrity

Cable systems utilize a wide range of RF frequencies, including those used by other radio systems for open-air transmission. To ensure a lack of interference to those licensed users of the spectrum, the FCC has enacted and enforces strict signal leakage standards.

Must-Carry/Retransmission Consent Issues

In addition to the digital transition issues discussed earlier are a host of other issues related to cable carriage of over-air broadcast signals, including which

signals must be carried, where they are placed in the RF spectrum, what quality is required, whether format conversions are allowed or required, and what is fair compensation for use of the programming. A related issue is cable's obligation for carriage of over-air broadcast signals relative to that of DBS operators.

Public, Educational, and Governmental (PEG) Access Issues

Franchised cable television operators currently negotiate with franchisors (which may be local cities, states, or a combination thereof) regarding access by various public entities to channels on the cable system. The usual use of such channels is for distribution of local programming (such as transmission of city council meetings to local residents), but sometimes facilities are demanded for internal communications needs, such as two-way data communications between civic buildings.

Access by Cable Operators to the Utility Poles and the Public Rights-of-Way

A cable franchise or OVS agreement provides cable operators with access to the public rights-of-way. Overhead facilities within those rights-of-way, however, are usually controlled by third parties, such as incumbent telephone operators, who may not wish to encourage construction of networks that could compete with their core businesses. One interest of regulators, therefore, is in ensuring that access to existing pole lines is granted on a compensatory, but reasonable, basis to cable operators.

A related issue is access to ground-level and below-ground rights-of-way. Regulators are concerned about the appearance of neighborhoods and the integrity of their streets, while operators need prompt access under financially reasonable conditions to build networks. It is essential that regulators understand what is technically feasible in addressing these issues.

1.3 The Development of the Cable Television Industry and Its Services

Cable television systems were not originally intended to be general-purpose communications mechanisms. Only recently has cable's role expanded to include nonanalog video services. Even then, these services are designed as much as possible around the performance required for analog video transport — for instance, they get the benefit of the high carrier-to-noise ratios required for analog video transport but are also generally constrained to operate within the 6-MHz channelization scheme that is a byproduct of the over-air frequency assignments.

The earliest cable systems were designed solely to be able to transport a common spectrum of analog television signals from a central point, called the

headend, to each subscriber's home, where they would connect directly with the television receivers (in the parlance of cable operators, this is referred to as the *downstream* direction). The systems and receivers were designed for signals conforming to the U.S. television standard, called *NTSC*, after the organization that created it in 1941, the National Television Systems Committee. This black-and-white television standard was modified in 1953 to provide compatible color information to color receivers and again in 1984 to add compatible stereo sound.

The original business case was based on the delivery of broadcast television to areas where they were not receivable with reasonable off-air antennas. In rural areas this was usually due to the lack of available signals, whereas in cities multiple reflections from tall buildings created many reflections ("ghosts"). These early systems were called *community antenna television* (CATV) and consisted of sometimes-elaborate off-air antennas constructed where required for good reception, coupled to broadband distribution systems that carried the signals to customers' homes.[4] Although other transmission lines were used in some early systems, coaxial cable soon dominated. Because of the relatively high loss of then-available cable, the technical progress of the early industry was paced by improvements in the technical properties and reliability of repeater amplifiers, cable loss, and such accessories as subscriber taps and connectors.

By the late 1960s, nearly all the areas of the United States that could benefit from a CATV system had been served. In the mid-1970s, the launch of satellite delivery of signals to cable systems changed the business model from simply extending the reach of over-air terrestrial broadcasters to dramatically increasing program choices for customers and actively competing with the broadcasters for audience share. The seminal event was the cablecast by the pioneering movie channel Home Box Office (HBO) of the October 1975 Ali–Frazier heavyweight boxing match from Manila.

Three new categories of cable-exclusive channels came into existence: (1) "superstations," local over-air broadcasters using satellite transmission to gain a nationwide audience; (2) specialized channels for news, sports, weather, education, shopping, and so on; and (3) movie channels, such as HBO, which sparked new excitement in the business. With this new programming, cable was able to tap into many new revenue streams and to offer television programming tailored to customers' wishes. Within a few years, customer video offerings typically included the following.

- A basic level of service typically including local broadcast stations, franchise-mandated local access channels, and perhaps a few additional channels
- One or more enhanced tiers of satellite-delivered specialized channels for additional monthly fees
- A choice of commercial-free movie and special-event channels, for an additional monthly fee per channel

- The ability to purchase access to individual showings of movies or special events, known as pay-per-view, or PPV

New, nonsubscriber, revenue streams came from sales of commercials inserted in satellite-delivered specialized channels and from a share of the sales revenues earned from "home shopping" channels that were usually carried as part of the basic service.

The availability of many possible television channels, and the possible revenue to be gained by carrying them, led to a continuous expansion in the bandwidth of cable systems. The earliest systems were limited to all or part of the 12 VHF broadcast channels (2–13). The first expansion utilized the spectrum between the top of the FM band at 108 MHz and the bottom of channel 7 at 174 MHz. After that, channels were added above channel 13. Over the years the maximum frequency has increased in steps from 216, through 270, 300, 330, 400, 450, and 550, to 750 and sometimes 870 MHz. At least two experimental systems were built with a maximum frequency of 1,000 MHz. Appendix A lists the now-standardized channel numbers and frequency ranges. The need to periodically upgrade distribution systems is one factor that makes cable television a capital-intensive business.

Delivering multiple levels of programming also required the development of various methodologies for selectively controlling access to individual channels, groups of channels, and individual programs. The need to provide and periodically upgrade these terminal devices is a second factor that drives cable's capital requirements.

Since the early 1990s, cable operators have actively tried to capitalize on their network investment to deliver additional services, both video and nonvideo. While there are exceptions, the general characteristics of these new services are that they:

- Require two-way communication between customers and the headend
- Are characterized by communications between the headend and individual customers
- Utilize digital transmission

One of the first additional services was the delivery of larger quantities of video that would fit in the existing or anticipated bandwidth of cable networks if transmitted in standard NTSC analog format. Digitally compressed video with quality comparable to analog video was developed at approximately the same time (and by overlapping groups of people) as high-definition television in the United States.[5] It has since become the enabling technology for DBS, as well as the emerging technology of the U.S. over-air broadcast stations. What it did for cable television is to allow six to twelve digital programs to fit in the same bandwidth as a single analog program, coupled with improved encryption

security. This, in turn, enabled a new level of service, known as *near video on demand* (NVOD), whereby the most popular movies were offered with frequent start times (typically 15-minute to half-hour intervals).

Converting cable networks to handle two-way communications allowed the best use of this increased program capacity by offering an easy means to purchase individual events using upgraded set-top boxes that included upstream data transmitters. Although coaxial amplifiers had been designed to allow conversion to two-way communications since the 1970s, the upstream modules were seldom installed, lacking an incremental revenue stream to support the hardware investment.

Conversion to two-way in large systems was, however, problematic because of the accumulated noise buildup in the upstream direction in large all-coaxial networks with cascades of 20 or more series-connected amplifiers. Beginning in about 1990, however, the use of linear fiber-optic technology allowed much shorter cascades and broke large systems into many smaller systems, independently fed by fiber. This not only made two-way communications practical, but had the incidental advantage that the programming fed to each coaxial subnetwork could be customized to the customers fed from that section of plant.

This combination of developments, driven initially by revenue opportunities in delivering more video choice, led to the offering of additional services that depended on reliable two-way communications and network segmentation. Meeting the requirements for these services has lead to new ways of assessing systems. In particular, each of these services depends on *per-user bandwidth* rather than *total system bandwidth* as a primary measure of system capabilities.

- High-speed data and Internet access in particular have become a fast-growing business segment for the cable industry. Cable companies provide two-way data paths between each user and the Internet, but they also may lease modems and sell retail services, such as Web hosting and e-mail. As noted earlier, the industry has been very successful in this market, outselling their competitors by a significant margin.
- True video-on-demand (VOD) differs from NVOD in that users can select movies from a library, control start times, and, during playback, have VCR-like control — pause, rewind, fast forward, and stop. It also differs from any of the previously listed video services in that a user-selected video event is sent privately to a single customer and thus requires dedicated bandwidth for the duration of the playback. At the end of 2002, VOD had moved beyond trials and wide-scale deployments were just beginning. One attraction to VOD is that it can't be duplicated by DBS, which can neither segment its audience sufficiently, nor support the required two-way communications.
- The move to offer voice telephone services presented the greatest technical challenge to cable operators. Although the bandwidth requirements are modest, meeting the BellCore reliability standard of 99.99% service availability was beyond the capabilities of most cable networks. Those

operators offering a primary-line service generally invested additional capital in "hardening" their networks, primarily through installation of more redundant field power supplies, but also committed to higher staffing levels.

Clearly, this is not an exhaustive list. Any service that can be modulated onto an RF carrier of reasonable bandwidth can potentially share a cable television network. In the future, we can expect to see e-commerce applications, Internet access from television receivers, interactive gaming, and a host of other applications.

In the remainder of this chapter we will discuss, at a high level, the evolving networks that supported this business model.

1.4 Cable Network Design

Since cable television originally was not a general-purpose communications mechanism but, rather, a specialized system for transmitting numerous analog television channels in a sealed system, the topology, or layout, of the network was customized for maximum cost efficiency in performing that function. The architecture that evolved before the introduction of fiber optics was called *tree-and-branch*.[6]

There are five major parts to a classical cable system: (1) the headend, (2) the trunk cable, (3) the distribution (or feeder) cable in the neighborhood, (4) the drop cable to the home and in-house wiring, and (5) the terminal equipment (originally as simple as the subscriber's television set but sometimes with an intervening converter in case the tuning range of the receiver was not sufficient).

The headend is the origination point for signals in the cable system. Signals are received there from off-air antennas, parabolic satellite antennas, terrestrial microwave links, or coaxial links from local sources. Sometimes programming is generated locally. Each channel is locally processed and converted, or modulated, to the proper television channel for distribution to customers. Then the channels are combined into a complete spectrum through a process known as *frequency-division-multiplexing*, or FDM, before insertion into the coaxial distribution system.

The distribution system has the task of delivering the complete FDM spectrum to each customer's home while meeting certain quality standards. At various points in cable's history these standards have been voluntary, have been set by various franchising agencies, or have been preempted and set by the Federal Communications Commission. They are also driven by the need to compete with various other television entertainment sources, including pre-recorded media. Currently, FCC rules govern the minimum acceptable performance, and are summarized in Chapter 15 (Table 15.1).

It is worth noting at this point that, unlike some other communications systems, cable operators generally measure wideband thermal noise levels separately from levels of discretely identifiable signals and distortion products. They do so because such noise sources arise from different physical mechanisms and affect picture quality in different ways. Thermal noise appears in analog

television pictures as "snow," while discrete signal interference appears as parallel patterns of lines with various spacings and angles. Most commonly, thermal noise is a function of the number and input levels of repeating amplifiers, whereas discrete interference can arise from intermodulation products or over-air signals that "leak" into the cable system or subscriber's television receivers.

Common design parameters for distribution systems (not including any degradation that may occur in terminal equipment) call for a carrier-to-thermal-noise ratio (C/N) of 48–49 decibels (dB) and carrier-to-discrete-signal-interference ratio of 53 dB. These target values are consistent with subjective data compiled in 1991 in a project sponsored by CableLabs. That work, supervised by Dr. Bronwen Jones, indicates "perceptible, but not annoying" thermal noise at a C/N of about 47–50 dB. The "slightly annoying" level was within about 1 dB of 41 dB.[7]

The most efficient coaxial design network for large systems evolved to consist of cascades of amplifiers of identical gain separated by lengths of cable whose loss just matched the amplifier's gain (although the loss through coaxial cable varies approximately as the square root of frequency, "equalizing networks" were used to make the loss flat across the entire FDM spectrum). Thus, the input and output levels of each amplifier were nominally identical. Under those circumstances, the carrier-to-noise and carrier-to-intermodulation-product ratios degraded as the logarithm of the number of cascaded amplifiers.

In a typical network, a relatively long "trunk" cascade of amplifiers, optimized for noise performance, would branch as required to feed shorter "distribution" legs, which usually contained only two or three cascaded amplifiers operated at relatively higher levels, with the intervening cable spans feeding strings of customer "taps," where a portion of the signals would be diverted to feed individual dwellings. As Figure 1.1 shows, the whole network resembled a tree with branches and sub-branches.

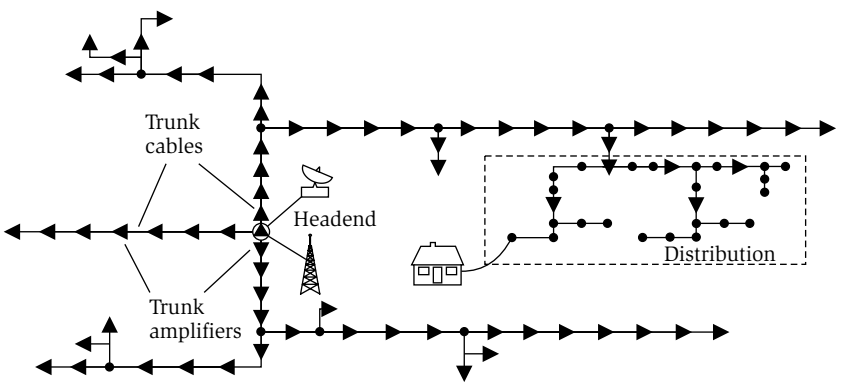


Figure 1.1 Tree-and-branch cable topology.

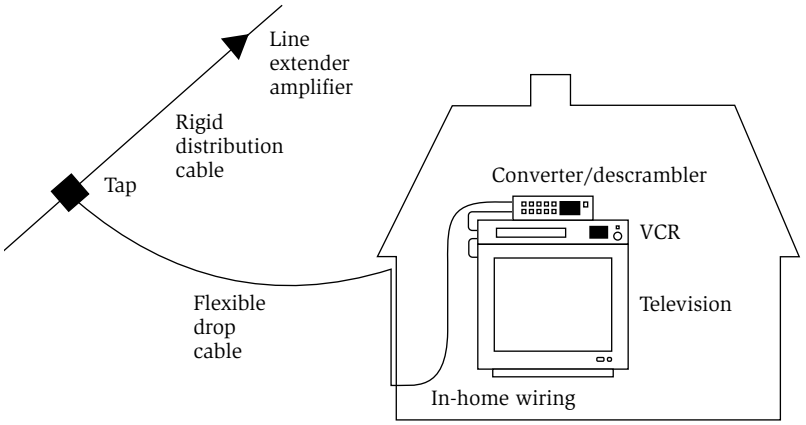


Figure 1.2 Terminal equipment and cable drop.

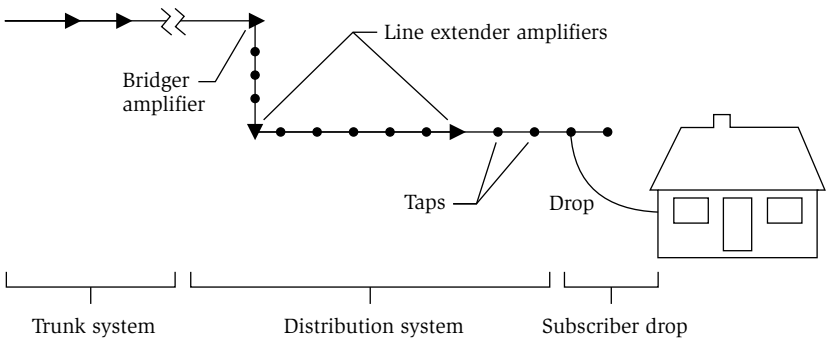


Figure 1.3 Distribution plant.

At each customer's house, a flexible cable was attached to the tap and routed to the terminal equipment, as shown in Figure 1.2. Figure 1.3 distinguishes between trunk, distribution, and drop.

1.5 Coaxial System Limitations

After some experimentation, the optimal design for coaxial trunk lines evolved to strings of amplifiers with gains of approximately 20 dB inter-

spersed with coaxial cables with the same loss at the maximum frequency being carried. In this configuration, each amplifier contributes an equal amount of noise (inversely proportional to input signal level) and an equal amount of intermodulation distortion products (proportional to some power of the output signal level and approximately the square of the number of signals being carried).

The development of cable television distribution technology has been driven primarily by the ever-rising downstream upper frequency limit, which in turn is driven by the need to carry increasing numbers of modulated signals. A secondary need has been for physically larger systems as smaller, formerly independent cable operators have consolidated into large regional operations. In response to these needs, manufacturers have responded with the following.

- Amplifiers with reduced noise figures, and thus reduced per-amplifier noise addition, for a fixed input level
- Amplifiers with increased bandwidths to handle the required frequency range
- Amplifiers with increased power-handling capability, and thus reduced intermodulation distortion, for a given output power level and number of channels
- Cables with reduced loss, both through improved dielectric materials and through the use of larger sizes of conductors with reduced ohmic losses

Though increased gain might be thought of as an improvement, because fewer amplifiers would be required for a given distance, higher-gain amplifiers would have to operate with either reduced input levels (and thus contribute more noise) or higher output levels (and thus contribute higher intermodulation distortion). Conversely, reduced gain per amplifier would require additional amplifiers, and thus drive costs and power consumption higher. A 20-dB net gain (plus or minus a few decibels) has proven a good compromise through many cycles of network evolution.

The possible improvements in component technology are limited. Cables cannot be made arbitrarily large, due to manufacturing and installation costs; amplifiers cannot have higher power-handling capabilities without driving up power consumption and associated heat dissipation; and noise figures are limited by physics and the practical manufacturing costs for devices that are mounted outside and thereby exposed to temperature extremes. The result is that, for a given technology and carried signal complement, there is a maximum possible length of trunk line that will allow any defined level of noise and distortion to be met. Furthermore, if the bandwidth of that trunk line is enlarged and more signals are added, the possible maximum length and amplifier cascade will decrease, because each amplifier will add more noise, distortion, or both.

One possible solution to this dilemma is the use of terrestrial microwave to cover at least a portion of the distance that would otherwise be covered by the trunk cable. The FCC established the cable antenna relay service (CARS) for this purpose. When used for trunking applications, the entire FDM spectrum is up-converted to a microwave frequency at the transmit end and then down-converted to the original frequency range at each receiving site, with the recovered signals inserted into a conventional coaxial distribution network, a technology known as *amplitude modulated link*, or AML.

Such AML trunking typically has noise and distortion performance comparable to about a 10-amplifier coaxial cascade yet can easily cover a distance of 10 or more miles, compared with a few miles for the equivalent coaxial trunk performance. Whether it is lower in cost depends on construction conditions. There are many locations where it is impractical to build coaxial lines but simple to build a microwave link. In other circumstances, multipath AML systems were built, simply to shorten cascades or to serve larger areas than were possible using only coaxial technology.

1.6 The Introduction of Fiber Optics

The major breakthrough in distribution system technology came with the development of practical, linear optical transport. Digitally (on–off) modulated optics had been in use for some time in the telecommunications industry, but in the early 1990s sufficient linearity and optical power levels were developed that it became practical to modulate the light level of distributed feedback (DFB) lasers with the composite FDM spectrum of signals that cable operators wished to transport. As with AML microwave, the performance of a single optical link was approximately equal to that of a 10-amplifier coaxial cascade. However, since optical fibers have a loss that is approximately 1% of the loss of coaxial trunk cable, the opportunity was immediately opened to build larger systems, shorter coaxial cascades, wider bandwidths, or some combination of those parameters.

The earliest application of linear optical transport in trunking was to shorten cascades so that bandwidth could be increased, as shown in Figure 1.4. For several years, the decreasing cost and improving performance of optical technology resulted in ever-smaller node-serving areas, shorter cascades, and increasing bandwidth. Today, off-the-shelf optical transmitters have downstream upper frequency limits of up to 870 MHz and the capability of handling 100+ analog video channels plus other signals with demodulated C/N of about 53 dB and carrier-to-intermodulation-product ratios of 65 dB.

More recent developments have included the development of high-powered transmitters (up to +17 dBm) and the use of the 1550-nm as well as 1310-nm optical wavelengths. Using 1550 nm enables operators to take advantage of fiber's lower loss at that wavelength, the use of optical amplifiers, and access to wavelength division multiplexing (WDM), which allows multiple optical

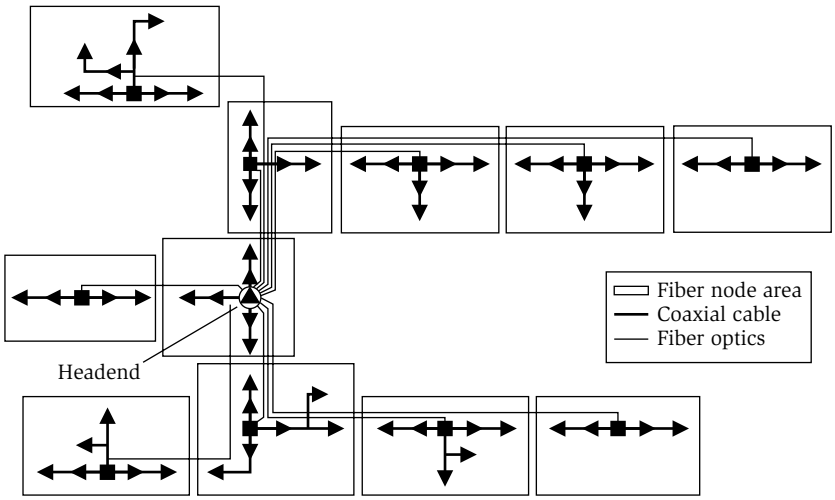


Figure 1.4 Cable system with hybrid fiber-coax topology.

signals to share use of a single fiber. Unamplified links in excess of 60 km and amplified links in excess of 200 km are now possible.

A side effect of the creation of smaller node-serving areas is the ability of operators to deliver an independent set of signals to each node, as discussed earlier. A second effect is that upstream signal transport becomes more durable, because the noise "funneling" from downstream amplifiers is reduced and the probability of ingress from external signals is reduced. These characteristics in turn allow practical delivery of subscriber-specific signals, such as are required for telephone, Internet access, and on-demand video services — all services for which communications reliability combined with the bandwidth per subscriber, rather than gross bandwidth, is of paramount importance.

1.7 High-Level Architecture Changes

Through a continuing process of industry consolidation, individual cable systems have grown steadily larger. This, combined with decreasing node sizes, has made it simply impractical to serve every node directly from the headend in a simple "star" network. Not only is the cost of running multiple, dedicated fibers to each node very high, the huge optical cables that would be required in a major metropolitan system would create major single points of failure.

As a result, operators have developed *distributed* networks, in which at least some of the headend functionality is moved to multiple hubs. Typically, the

transport between headend and each hub is entirely optical and redundant, with route-diverse transport for all critical signals. High levels of multiplexing, whether at the baseband digital level, through wavelength division multiplexing, or a combination of both, reduces the required fiber counts between major facilities. In the largest systems, the core headend signal-processing requirements are sometimes duplicated at two locations, so the entire headend is redundant.

1.8 Video Signal Security and Selective Service Delivery Issues

The creation of multiple marketable levels of television programming required the development of means to selectively control access to individual channels or groups of channels. The two earliest of these were known as *negative traps* and *positive traps*. Negative filters (or traps) were simply band-stop filters that *blocked* individual channels or groups of channels by removing a sufficient amount of a signal or group of signals to render the recovered pictures unwatchable; positive traps were narrow bandstop filters used to *enable* reception of individual channels by removing interfering signals inserted at the headend without removing a significant amount of the sideband information.

Traps, however, had many drawbacks: Negative traps damaged adjacent channels to a greater or lesser degree, and positive traps inevitably removed enough sideband information to noticeably degrade reception. These problems, as well as the inevitable insertion loss, increased as the upper bandwidth of systems increased. Additionally, traps offered only moderate security and could easily be removed by customers, modified, or simply duplicated, as required. Finally, they required physical access to install or remove, so modifying a subscriber's receivable channels required a trip to the home or nearest tap.

The issue of degraded reception was addressed through the development of "descrambling" set-top boxes. Signals to be protected were modified at the headend so as to be unwatchable on standard television receivers. The descrambler reversed the scrambling process for whichever channels were programmed into its local memory. In the United States, the most common technique was to eliminate or modify the synchronization pulses that are a crucial part of a modulated NTSC signal, although more sophisticated techniques (such as shuffling the order in which lines of a picture were transmitted) were employed in European systems.

Although many preprogrammed descrambling converters were deployed, they had the limitation that changes to the list of channels being descrambled still required physical changes to the converter. As a side effect, any box not recovered from an ex-customer represented a potential means by which someone else could access premium programming without paying for it.

A second generation of descrambling converters solved the problem of changes to enabled channels by adding *addressability*. An addressable descrambling converter incorporates the ability to download descrambling information

sent from the headend in real time. This not only reduced the cost of providing services sold on a monthly basis, but enabled the provision of PPV, where access needed to be enabled for just the duration of a single event.

Given market studies that showed that the purchase rates for pay-per-view events were strongly influenced by the ease with which a customer could access them, combined with the cost of staffing customer service centers to handle phoned-in orders, the industry began to deploy boxes that included upstream transmitters so that customers could order programming electronically using their remote controls. This was one of the factors driving the deployment of two-way upgrades to RF amplifiers.

Unfortunately, due to the need to control costs, the scrambling methodologies utilized for securing analog television programming were not particularly sophisticated, nor were the boxes "hardened" against modification, with the result that many were modified to enable reception of more channels than originally intended, and a black market industry in both modified and new "pirate" boxes developed.

The advent of digitally compressed television, as discussed earlier, allowed operators to carry many more channels in the same bandwidth. Due to the quadrature amplitude modulation (QAM) used, the signals were also more durable and could be transmitted through channels with somewhat lower fidelity than analog video channels (though still much greater than for baseband digital signals).

While much more secure, the digital boxes also require significant internal computational ability and digital memory capacity. Thus, they are more expensive than comparable analog addressable converters. Nevertheless, the advantages of digital are such that they are rapidly replacing analog techniques for security and control of video programming. They are almost universally two-way and include both a primary receive channel and a second channel for simultaneous reception of control data from the headend. This has provided a durable and flexible platform for interactive services, such as VOD, interactive gaming, and e-commerce, which require low-latency, high-bandwidth, two-way communications between the headend and individual customers.

1.9 Consumer Equipment Interface Issues

Except to the extent blocked by filters, downstream cable television signals are transmitted throughout customers' homes so that television receivers, set-top boxes, data modems, and other terminal equipment can be connected to any cable outlet. For the same reason, signals generated from anywhere in the home that are within the upstream transmission band (extending roughly from 5 to 40 MHz) will be carried toward the headend. Though this arrangement provides a lot of installation flexibility, it has also created several layers of interface issues for both system operators and consumers.

1.9.1 Shielding Integrity

As discussed earlier, cable television systems transmit signals on frequencies that are used for terrestrial transmission by other entities. This is possible because cable systems are well-shielded closed systems. The weakest link in this shielding is the wiring in customers' homes and the terminal devices connected within the home. By regulation, the wiring within each home is the property of the homeowner and may be installed or modified by them at will. The requirement on the cable operator is to monitor the entire plant for egress and, if owner-controlled wiring is leaking, to work with customers to remedy the problem or, in the ultimate, to disconnect service until it is remedied.

While egress is a legal problem, ingress is a greater threat to system operation. Generally, ingressing signals cause transmission problems before egressing signals exceed legal limits.

Defects in shielding can occur due to the use of poor-quality cable, loose or poorly installed connectors, inadequately shielded components (such as splitters or switches), or poorly shielded television receivers connected to outlets. Unfortunately, the regulations covering the shielding integrity of consumer electronics devices, including television receivers, are not adequate to ensure adequate system egress and ingress performance.

1.9.2 Directly Connected Television Receivers

Almost universally, cable operators deliver their lowest level of video service in analog, unscrambled form, so subscribers need only connect a television receiver with adequate tuning range directly to a broadband outlet in their home. Such "whole-house" reception capability is then marketed as a superior alternative to, for instance, DBS reception, in which a set-top box is required for each television set. In addition to shielding issues, however, connecting television receivers directly to cable outlets raises a number of other interface issues.

Received Signal Quality

Television receivers are designed for connection to antennas. In that environment, a primary requirement is adequate low-signal performance. Generally, fewer than 20 signals are present in the antenna downlead and adjacent channels are not present. By contrast, cable systems deliver roughly 100 channels on contiguous channels at relatively high signal levels. This requires receivers to have better signal-handling characteristics and adjacent-channel rejection filters.

Additionally, devices within the customer's home or the homes of his neighbors may generate upstream signals at levels more than 100,000 times higher than received signals. Due to finite system isolations, these signals will also be present at the input terminals of the receiver, so it must be designed to adequately reject them.

Finally, television receivers generate various internal signals used in signal-processing steps. To the extent that these are present at the input terminals and fall in the upstream band, they degrade two-way system performance. Thus, directly connected receivers need to be designed with high internal isolation.

1.9.3 Functionality Limitations Due to Set-Top Boxes

The imposition of a set-top box between the cable outlet and a subscriber's receiver, while it enables additional services to be delivered, affects functionality. For example, if a subscriber has a video cassette recorder (VCR) connected to the receiver, then the single-tuner set-top box will block the ability to record one channel while viewing another channel. Moving the VCR to a position ahead of the set-top box restores the ability for independent recording but prevents the user from recording any program that is scrambled. Similar difficulties arise when used with receivers capable of picture-in-picture display or when doing timed, multi-channel recording. Ultimately, all these difficulties arise from the fact that the set-top box has its own tuner, which does not operate under the control of the receiver.

1.9.4 Subscriber-Owned Terminal Equipment

Clearly, one answer is for the functionality of the set-top box to move within the receiver. By placing the descrambling process behind a tuner that is under the control of the user, most of the limitations caused by the use of external tuner/descramblers disappear. Despite the efforts described earlier, this is still just a future possibility in the television world. By contrast, a robust standardization effort has made it possible in the world of high-speed data services, with the result that cable modems are low in cost and are becoming readily available in retail electronics outlets.

1.10 Summary

In this chapter we have presented a brief overview of cable television's development and current practice in the United States. Starting with a few comments on cable's history, we have reviewed the structure of the cable industry and described the fundamentals of the technology on which it rides.

The industry began as a relay service for distant over-air broadcast signals and, while in that phase, developed the basic coaxial technologies and components required. The addition of independent sources of satellite-delivered programming changed the nature of the broadband business and drove the development of ever-wider-bandwidth systems and associated development of technology to selectively control access to programming.

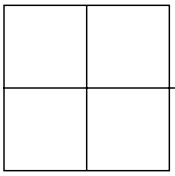
The use of linear fiber-optic technology was driven by a need to expand system size and bandwidth beyond what was possible using all-coaxial

technology, but this brought with it practical two-way transmission and the ability to achieve much higher per-customer bandwidths through the delivery of customized signals to each system node. This in turn enabled the launch of two-way services, such as high-speed Internet access and voice telephony.

Finally, the development of digital video compression has multiplied 10-fold the number of channels that can be transmitted through a given amount of bandwidth. This increased information capacity, combined with the features of digital video set-top boxes, has solved a security problem and also opened the door to the provision of sophisticated interactive services. The rest of this book will fill in the details. Now it's time to roll up our sleeves and dive in!

Endnotes

1. *Cable Television Developments*. Washington, DC: NCTA, summer 1999.
2. For example, draft *EIA Interim Standard IS23* (1994) covered RF characteristics, *EIA Standard EIA-105* (1998) defined an interface that could be used with either analog or digital decoder modules, and *EIA Standard EIA-679* (1998) defined a system whereby digital decryption security could be separated from the receiver.
3. "Time to Plug and Play — Cable, CE Industries Reach TV Accord," *Multichannel News*, December 23, 2002.
4. The early history of CATV and the people who played key roles in its development has been well documented in: Archer Taylor, *History Between Their Ears — Recollections of Pioneer CATV Engineers*, Denver. The Cable Center, 2000.
5. The nontechnical story of the development of HDTV, with reference to the development of digitally compressed video in general, is well told in: Joel Brinkley, *Defining Vision — The Battle for the Future of Television*. New York: Harcourt Brace, 1997.
6. Walter S. Ciciora, *Cable Television in the United States, An Overview*. Louisville, CO: CableLabs, 1995.
7. Bronwen Lindsay Jones and James A. Turner, *Subjective Assessment of Cable Impairments on Television Picture Quality*. Louisville, CO: CableLabs, June 1992.



Chapter 2
Analog Television

2.1 Introduction

One of the most successful technologies of our time is based upon the standard created by the National Television Systems Committee (NTSC). Like the U.S. Constitution, it seems to have anticipated much of its future and adaptively followed the evolution of its world. Starting with vacuum tubes and progressing to integrated circuits, analog television has survived and prospered. Though much of the current excitement centers on digital television, it is reasonable to assume that analog television will be with us for many more years. Even after digital receivers significantly penetrate households, they will serve the principal viewing room while older analog receivers will be found throughout the rest of the residence.

2.2 Baseband Monochrome Video and Monaural Audio

The technology of analog television is the foundation of the cable industry. An understanding of analog television is key to fully appreciating the rest of the subject matter of this book. Understanding the television signal is a basic requirement for designing systems to deliver it to subscribers.

2.2.1 Television Briefly

The Federal Communications Commission (FCC) adopted standards for commercial television broadcasting in the United States in 1941.[1] It is an analog system, in which the picture is transmitted via vestigial-sideband modulation of the visual carrier and the sound is frequency modulated on a separate carrier. In 1954, the NTSC "compatibly" extended the system to include color information by increasing the utilization of the 6-MHz spectrum occupied by the television channel while allowing continued use of essentially all older receivers with little performance degradation. In 1976, the FCC reserved line 21 for the closed captioning system, and in 1983 a teletext service was authorized. In 1984, the NTSC system was again compatibly extended to include stereo sound. Most

recently, in June 1996, the FCC authorized additional ancillary data services in analog television. These technologies will be considered in this chapter.

The NTSC television system is termed *analog* because the signals representing the picture and sound information can take on *any* value between the minimum and maximum limits. In particular, the strength of the visual signal is inversely related to the picture brightness, with black portions of the picture having the most transmitted power. This negative modulation of the video information was found to reduce the impact of Gaussian white noise on the displayed image. Periodic pulses are included at powers greater than those used to represent black areas in the picture. These pulses provide the timing information required to synchronize the transmitter and the receiver, allowing the picture to be painted correctly on the screen.

Understanding television technology is aided by some appreciation of its history.

2.2.2 Scanning

A picture can be regarded as the composite of many points of light called *picture elements*, or *pixels*. The information for each of these points of light must be derived from the object or scene to be televised, organized in a logical manner, conveyed to the receive site, and then displayed as an image. Since the number of points of light needed to make up a picture is huge, it is not practical to convey to each of them via its own transmission channel. Scanning is a simplifying process that allows a single transmission channel.[2]

The process of scanning was first used on a large scale to transmit still images for use in newspapers (see Figure 2.1). A photograph is placed on a rotating drum and a light-sensitive device is focused on it. As the drum rotates, the light-sensitive device follows a path across the image. With each revolution of the drum, the light-sensitive device is advanced down the length of the rotating drum. This creates a helical scan that follows the path of an imaginary

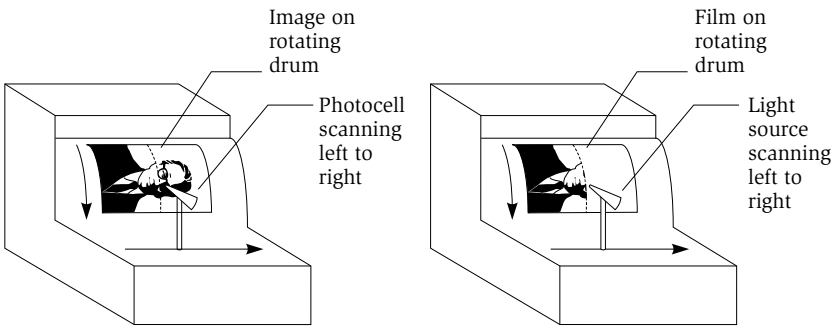


Figure 2.1 Scanning of photograph on rotating drum for transmission to receiver with light source and photosensitive film.

coil spring across the photograph on the scanning drum. The electrical output of the light-sensitive device is proportional to the reflectance of the image, bright areas yielding a larger electrical signal than dark areas. At the receive location, the process is reversed. A light-sensitive film is attached to a rotating drum, and a light source scans the same helical path across the film. High electrical input results in a bright light and a full exposure, or blackening of the film. Conversely, low electrical input yields a lower exposure, or a more transparent area on the film. In this way, a negative image is produced. Standard photographic techniques yield a comprehensible image. It is important that the rotating drums at the transmit site and the receive site be synchronized. They must rotate at the same speed, and the start of scanning must occur simultaneously in both locations. Likewise, the rate of advance of the helical path down the drum must be the same in both locations. If this synchronization does not occur, then geometric distortions are introduced in the reproduced image. Modest distortions involve stretching or compressing of the reproduced image as if it were printed on a rubber sheet and then the rubber sheet distorted. More destructive distortions are possible if the speeds are very different in the two locations.

Although scanning works well when the receive site includes a photosensitive film that can accept and retain the image, the problem becomes a bit more complex when the receiver is the human eye. The human eye exhibits a phenomenon termed *persistence of vision* that allows it to continue "seeing" a light for a brief time after the light is removed. The application of the scanning principle to direct human vision requires that the scanning process be executed repetitively and quickly enough for the eye to fuse all the scan lines into an image.

In 1883, the German engineer Paul Nipkow invented a scanning method consisting of a rotating mechanical disk with appropriately located holes arranged in a spiral (see Figure 2.2). As the disk rotates, only one hole at a time is over the image to be transmitted (see Figure 2.3). As one of the holes completes a left-to-right scan, another hole commences scanning from the left slightly below the path of the previous hole. When the bottom hole completes its scan, the scanning disk will have made a full revolution, and the top of the image is again scanned. Nipkow was granted a German patent but did not actually build the system. In 1923, John Logie Baird of the United Kingdom built a crude system based on the Nipkow scanning disk. It was clumsy and produced poor pictures, but it confirmed the principle of scanning.

Television did not become practical until electronic display tubes and pickup tubes were introduced. In 1923, Vladimir Zworykin, a Russian immigrant who became a U.S. citizen, applied for a patent for a completely electronic television system. The interest in television was intense at that time, and many inventors were active in the field. Space does not allow a discussion of the numerous contributions and the interesting patent fights among inventors. This fascinating story is described elsewhere in detail and is recommended reading.[3]

In 1929, Zworykin applied for a patent that made the monochrome (black-and-white) picture tube practical (see Figure 2.4). An *electron gun* launches electrons

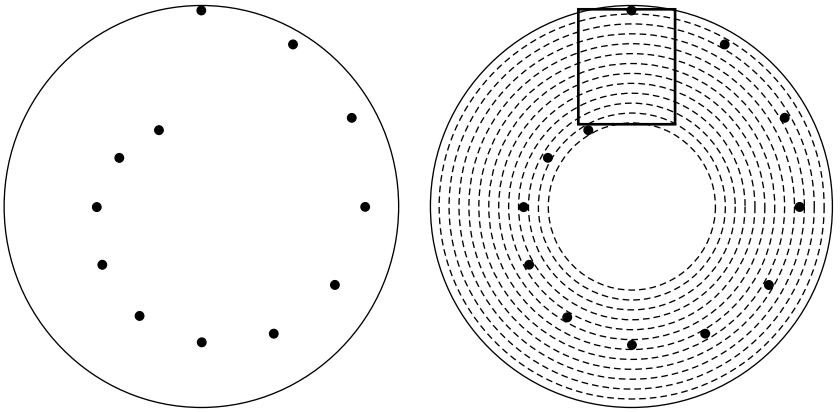


Figure 2.2 Nipkow disk operation. (a) Nipkow disk with a spiral of apertures. (b) Nipkow disk's scanning trajectory through the image plane.

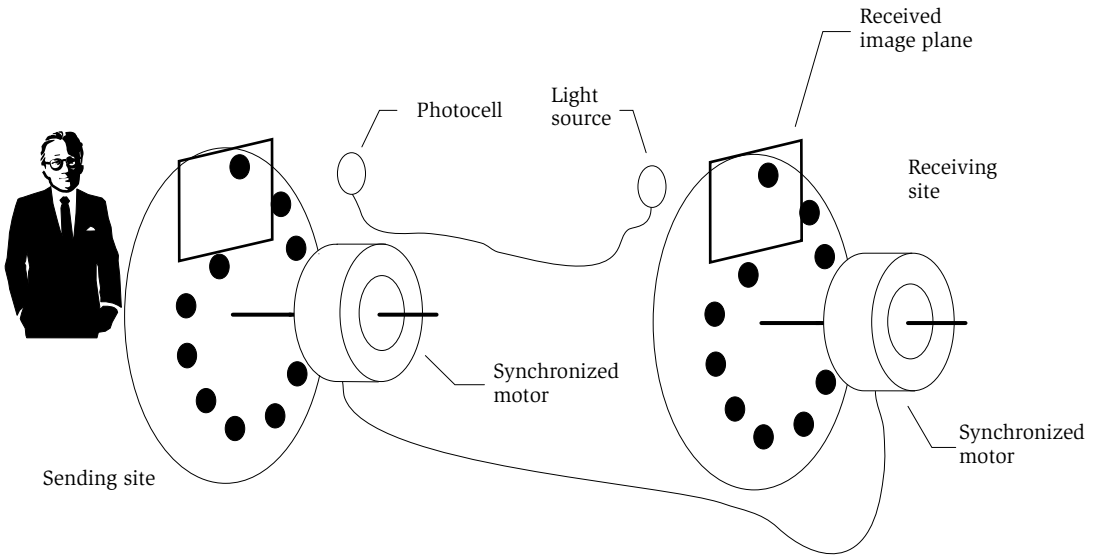


Figure 2.3 Only one hole at a time is exposed to the image on the sending side and the light source on the receiving side.

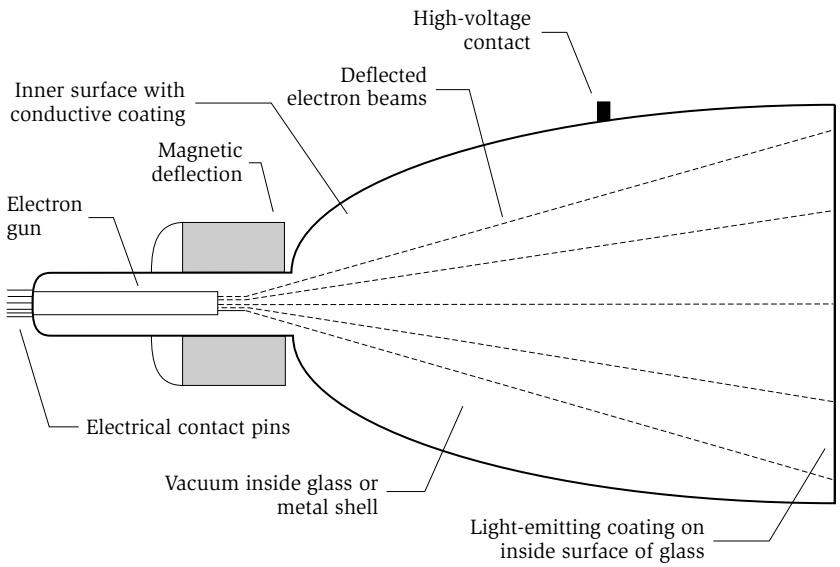


Figure 2.4 Electronic television is based on the kinescope, or cathode ray tube.

past a control grid that regulates the strength of the electron beam. The beam then passes through focusing elements that concentrate the electron beam into a small spot on a glass faceplate that is covered with a fluorescent coating on its inside surface. When the electron beam strikes the fluorescent coating, light is emitted in proportion to the strength of the electron beam. Magnetic coils on the neck of the tube deflect the electron beam vertically and horizontally to implement scanning. The scanned rectangle of light on the face of the picture tube is called a *raster*. The picture tube is also called a *kinescope* because it was meant to display moving or kinetic images. Yet another name for the kinescope is the cathode ray tube (CRT), since the part of the electron gun that emits the electrons is called a *cathode*.

In October 1931, Zworykin applied for a patent on a practical electronic television camera tube that he called the *iconoscope* (see Figure 2.5). As with the kinescope, an electron gun forms an electron beam that is scanned across a surface by magnetic deflection coils. In the case of the iconoscope, the surface is coated with a light-sensitive material. Light focused on the sensitive material causes it to develop an electrical charge in proportion to the strength of the light. When the electron beam strikes the light-sensitive surface, it interacts with the surface in proportion to the accumulated charge. A signal plate on the back side of the light-sensitive surface produces an electrical signal proportional to the light that fell on the surface. Again, many other inventors were working on similar devices, and more patent struggles occurred.

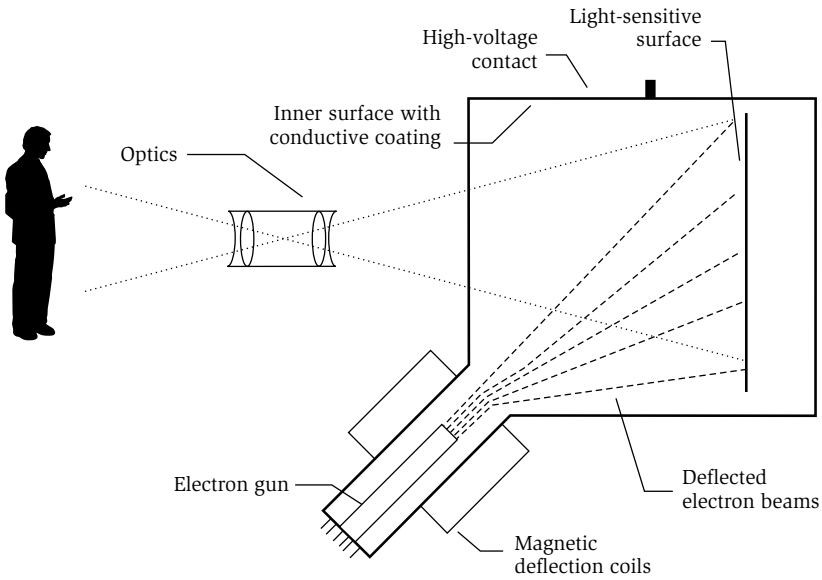


Figure 2.5 Electronic television is also based on Zworykin's iconoscope.

After a scan line is completed, it is necessary to move the electron beam back to the left side of the screen.[4] Ideally, this *retrace* would be done instantaneously, since no picture information is transmitted during the backward motion. Both the scanning drum of Figure 2.1 and the scanning disk of Figure 2.2 did not require retrace. However, electronic deflection is accomplished with magnetic fields, and the magnetic fields store energy. Deflection circuits require time to change the magnitude and direction of the energy stored in the magnetic deflection fields. As a result, it is not possible to instantaneously move the electron beam from the right side of the screen to the left or from the bottom to the top. During the time the electron beam is in retrace, it must be turned off (a process called *blanking*). The period during which the beam is turned off in horizontal retrace is called the *horizontal blanking interval* (HBI), while the much-longer period required to move from bottom to top is the *vertical blanking interval* (VBI). The combination of the two blanking intervals constitutes around 25% of the total time and is the price paid for using cost-effective magnetic deflection. It is interesting to note that modern liquid crystal display (LCD) panels do not require retrace time.

Thomas Edison is credited with developing practical *moving pictures*. Moving pictures depend upon a rapid display of a succession of still pictures. In modern motion picture practice, 24 pictures are each displayed twice per second, but in early movie history the rate was reduced, leading to noticeable "flickers."

The human response to flicker is very complex and is related to the color, size, and brightness of the image. A small increase in the brightness, or a decrease in the repetition rate, results in a much higher degree of noticeable flicker. The minimum rate that is not objectionable is called the *flicker fusion frequency*. A flickering light source at frequencies above that value appears as continuous illumination. For example, fluorescent tubes operated at 60 Hz (and thus extinguished 120 times per second) appear to be continuously "on" because they are operated above the flicker fusion frequency. Monochrome television developers settled on 30 complete images per second in the United States and 25 images per second in Europe. Because of the interlace technique, described later, 60 half-images (or 50 in Europe) are displayed per second. These choices were originally tied to the power line frequency causing any unfiltered power line "hum" that got into the picture to be stationary on the screen. (It was determined that slowly moving patterns of interference were much more objectionable than stationary patterns.) In the European systems, the brightness level at which television flicker is not objectionable is only about 30% of that in the United States. Many viewers accustomed to 60-pictures-per-second television find that the European systems flicker excessively at the brightness levels normally used.

Many of the specifications for the NTSC television system were based on the desire to minimize the cost of receivers using then-current technology and are no longer required. For example, it is now practical to keep the power line hum completely out of video circuits. Consequently, the need to relate image display rate to power line frequency is no longer significant. In Japan, for instance, 50-Hz power is used in the Osaka region and 60-Hz power in the Tokyo area, yet the same television system serves both markets.

The amount of information required to produce pictures of acceptable horizontal and vertical resolution at a 60-pictures-per-second repetition rate is huge. At current resolutions but with a repetition rate of 60 complete pictures per second (that is, without the interlace technique described later), twice the information would be required, and a baseband bandwidth of 8.4 MHz would be required for the video alone. As we shall see, the modulated bandwidth using simple amplitude modulation would be 16.8 MHz — almost three times the current broadcast bandwidth of 6 MHz. Clearly, ways of reducing the bandwidth had to be found that did not impair perceived picture quality.

One of the earliest signal compression techniques involved modifying the scanning system. Sixty pictures per second are required in television because picture tube technology could produce pictures bright enough to watch in daylight with the window drapes open without apparent flicker. Since experience with motion pictures confirmed that 24 complete images per second is an adequate information rate for motion rendition, a technique called *interlaced scanning* was developed whereby half of the scan lines are displayed in one "field" and the other half of the scan lines in another "field." Complete two-field "frames" are transmitted at a rate of 30 per second, providing better motion rendition and flicker frequency than movies yet requiring just 4.2 MHz of baseband bandwidth.

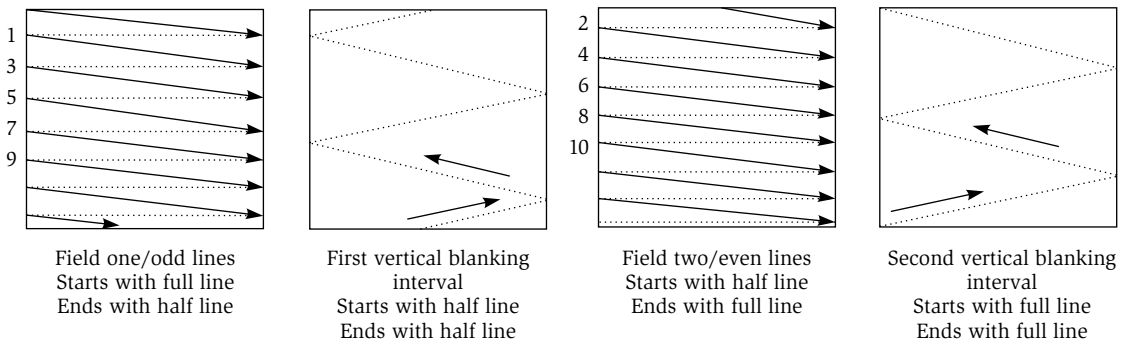


Figure 2.6 Interlaced scanning to reduce flicker while conserving bandwidth.

Interlaced scanning is illustrated in Figure 2.6. A 525-line raster includes two fields of 262.5 lines each. The odd fields start with a whole line and end with a half line, whereas the even fields start with a half line and end with a whole line.

Even though interlace scanning works well with most "natural" video (that is, camera images), interlace scanning can produce inferior results with "synthetic video," which is electronically generated and does not benefit from some of the averaging phenomena of practical cameras. For example, the electron beam in a television camera has a distribution of charge of roughly a Gaussian shape. As a result, it accepts and averages light from adjacent scan lines. This averaging causes a smoothing of the image. Synthetic video, especially when producing text, sharply defines the images of each line. This results in a brightly illuminated edge that is presented at 30 images per second, well below the flicker fusion frequency. This is unacceptable in computer displays that are bright and utilized at very close viewing range. As a consequence, nearly all computer displays use a simple top-to-bottom "progressive scan" technique. Since computer displays are driven by locally generated video, the higher bandwidth requirement of a progressive scan raster is not a problem.

2.2.3 Synchronization

As previously described, there are two fields in each frame of video.[5] Frames have 525 scan lines and repeat 30 times a second, resulting in 63.5 microseconds per horizontal scan line. The reciprocal of this yields the 15,750-Hz monochrome horizontal line frequency. Since fields occur at the rate of about 60 per second, a field lasts 16.67 milliseconds. (All of these values were slightly modified when color was added — see Section 2.4.1)

A horizontal synchronization pulse is required for every line and a vertical synchronization pulse is required for every field in order to display a stable

image. Synchronization (sync) signals are distinguished from the composite signal on the basis of amplitude. The voltage range for visible brightness variations extends from 12.5% to 75% of the unmodulated carrier voltage, leaving 25% for the sync pulses.

Figure 2.7(a) displays an example of a signal that might come from a television camera as it scans one line. At 75% of the maximum voltage, a blanking pulse is inserted, timed to occur during the retrace time. This is shown in Figure 2.7(b). The blanking pulse is nominally 10.9 microseconds, or 16.5% of the horizontal period, while active video occupies 52.6 microseconds. The blanking pulse is at the video signal's black level, keeping the picture tube from emitting light during retrace. The sync pulse is installed on top of the blanking pulse and is at a level referred to as "blacker than black" (see Figure 2.7(c)). The sync pulse is nominally 4.7 microseconds long. Figure 2.8 shows the relative levels of black, blanking, and sync pulses in the modulated envelope of the signal.

Figure 2.9 shows the horizontal sync pulses and video of Figure 2.8 plus additional sync pulses for the vertical blanking interval. Each field's VBI is 21 lines long, leaving 483 lines for active video. The first nine lines of the VBI are devoted to pulses associated with the vertical synchronization process. The fourth, fifth, and sixth lines in each field contain the vertical sync pulse. This pulse is serrated, creating pulse edges properly located for the horizontal synchronization to be maintained during the vertical retrace. The vertical sync pulses are called *broad pulses*, in comparison to the narrow horizontal sync pulses.

Simple resistor–capacitor *sync separator* circuits can be used to separate the vertical pulses from the horizontal pulses.[6] In the lower portion of Figure 2.10(a), capacitor C1 is charged through series resistor R1 and discharges through the input impedance of the following circuits. The narrow horizontal pulses will not build up a significant voltage because the charge bleeds off during the relatively long time between horizontal sync pulses. However, the broad pulses will build up significant charge. This is shown in more detail in Figure 2.10(b). Those who have studied such circuits will recognize this configuration as a simple low-pass filter, also called an *integrator*. A properly adjusted level-sensitive circuit following the resistor–capacitor network will provide a vertical timing pulse. This pulse is used to synchronize a vertical deflection oscillator, which in turn drives the vertical deflection magnetic coils on the neck of the picture tube. Six "equalizing pulses" are inserted before and after the broad pulses to equalize the charge buildup on the capacitor, preserving the precision of the vertical timing. This is necessary because one field ends with a half line and starts with a whole line, whereas the other field has the opposite configuration. Without this precaution, the timing between the two fields will be upset, and the rasters they produce will not be evenly interlaced. The resulting distortion is called *pairing* because adjacent lines from both fields pair up and blur each other's detail while leaving dark lines between them. Although modern digital sync separator circuits do not need the equalizing pulses, the original receivers using vacuum tubes had to rely on tailoring the signal to allow simple circuits.

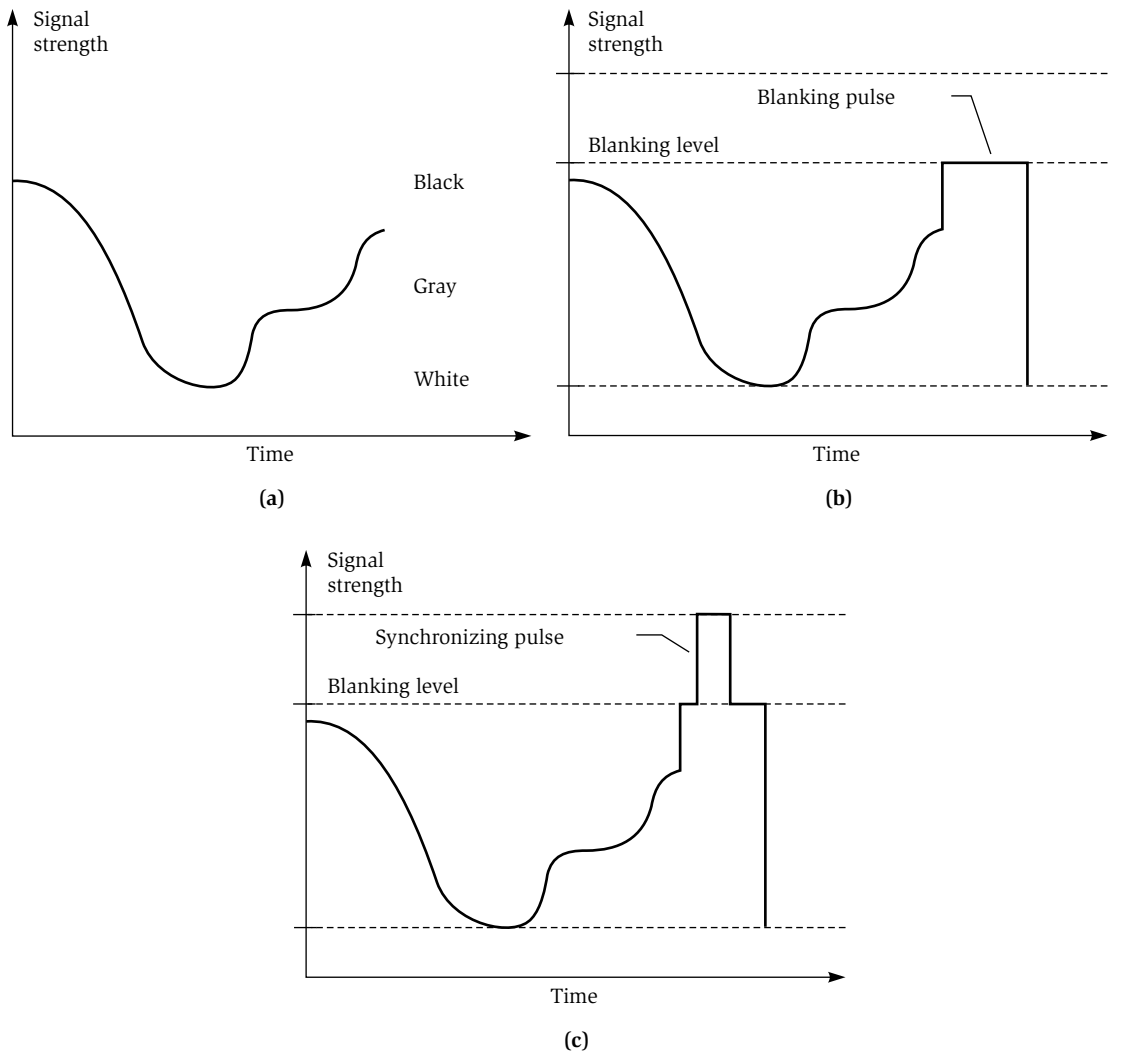


Figure 2.7 Construction of a video signal. (a) Video signal from television camera. (b) Blanking pulse added to avoid diagonal retrace lines. (c) Horizontal synchronization pulse added to facilitate left-to-right timing.

In the upper portion of Figure 2.10(a), a series capacitor C2 with a resistor R2 to ground will convert the edges of the pulses into positive- and negative-going spikes that can be detected by appropriate level-sensing circuits. Note that the serrations in the broad pulses allow a continuous flow of uninterrupted pulses at the correct timing for horizontal synchronization. When these pulses occur

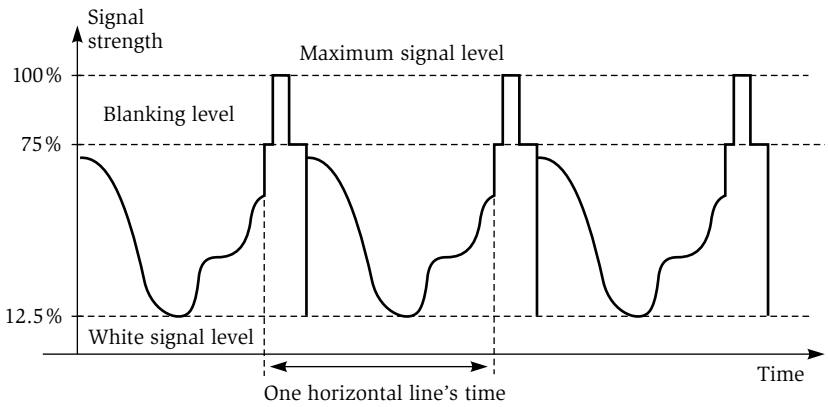


Figure 2.8 Composite monochrome television signal levels (not to scale).

halfway through a horizontal line, the horizontal circuits ignore them. They are effective only when they are near the correct time. Those who have studied such circuits will recognize this configuration as a simple high-pass filter, also called a *differentiator*.

By allocating significant signal resources (about 25% of the signal time and about 25% of the voltage range), rugged synchronization of the video signal was achieved.

2.2.4 Spectrum of Scanned Images

It is well understood that any repetitive signal can be decomposed into a sum of sine and cosine waves of appropriate amplitude, frequency, and phase.[7] There

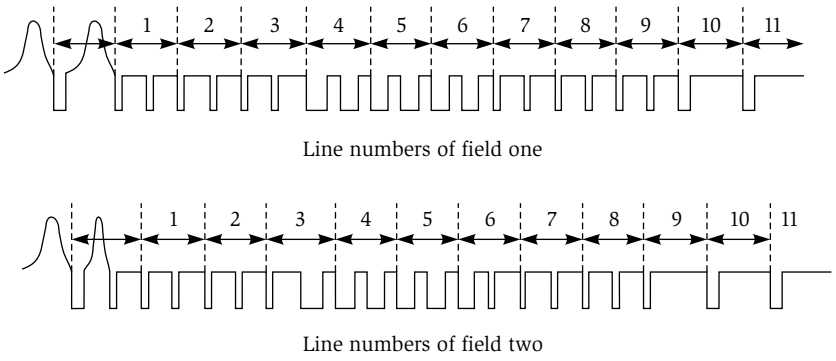


Figure 2.9 Vertical and horizontal synchronizing pulses of field one and field two. Note locations of half lines (not to scale).

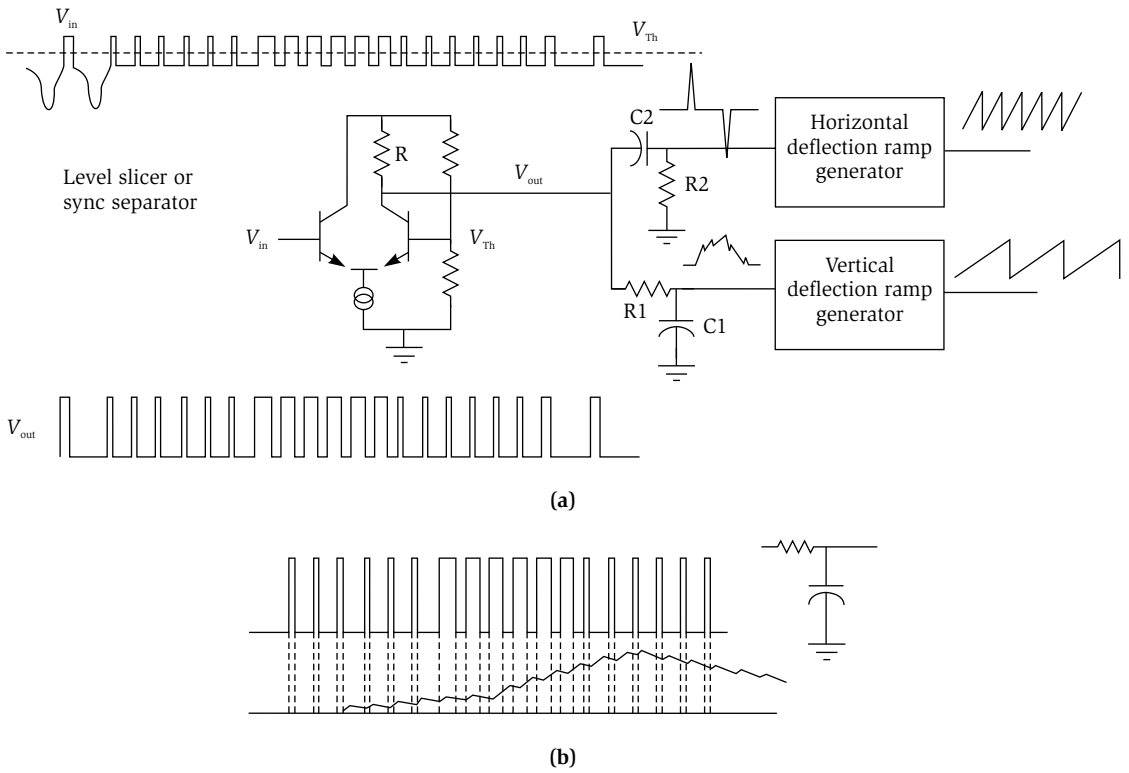


Figure 2.10 Receiver synchronization. (a) Sync separation and deflection ramp generation. (b) Vertical synchronization pulse generation.

are at least three ways we know this: (1) mathematical proof, (2) geometric construction, and (3) measurement with spectrum analyzers. The most rigorous method is the mathematical. Equations (2.1), (2.2), and (2.3) display the familiar Fourier series equations.

$$f(t) = a_0 + \sum_{n=1}^{\infty} [a_n \cos(n\omega_0 t) + b_n \sin(n\omega_0 t)] \qquad (t_0 < t < t_0 + T) \qquad (2.1)$$

Equation (2.1) is the trigonometric Fourier series representation of $f(t)$ over an interval $(t_0, t_0 + T)$.
The various constants a_n and b_n are given by

$$a_n = \frac{\int_{t_0}^{(t_0+T)} f(t) \cos(n\omega_0 t)\ dt}{\int_{t_0}^{(t_0+T)} \cos^2(n\omega_0 t)\ dt} \qquad (2.2)$$

and

$$b_n = \frac{\int_{t_0}^{(t_0+T)} f(t)\sin(n\omega_0 t)dt}{\int_{t_0}^{(t_0+T)} \sin^2(n\omega_0 t)dt} \tag{2.3}$$

We could derive these equations and apply them to scanned images, but that would duplicate what is already available in many other places. In particular, a paper published by Mertz and Gray in the *Bell System Technical Journal* in July 1934, titled "Theory of Scanning and Its Relation to the Transmitted Signal in Telephotography and Television," showed that periodic scanning of an image yields a spectrum with energy clustered around the horizontal scanning frequency, with large gaps in between. This pioneering work was done long before the availability of modern computers. Donald Fink did follow-up work and the results are included in the 1957 *Television Engineering Handbook*.[8] His work clearly showed that the picture's energy is clustered around harmonics of the horizontal scanning frequency, as shown in Figure 2.11.

2.2.5 Television Audio

Television sound is limited to 15 kHz because of the presence of strong spurious signals related to the monochrome horizontal scanning frequency of 15,750 Hz.

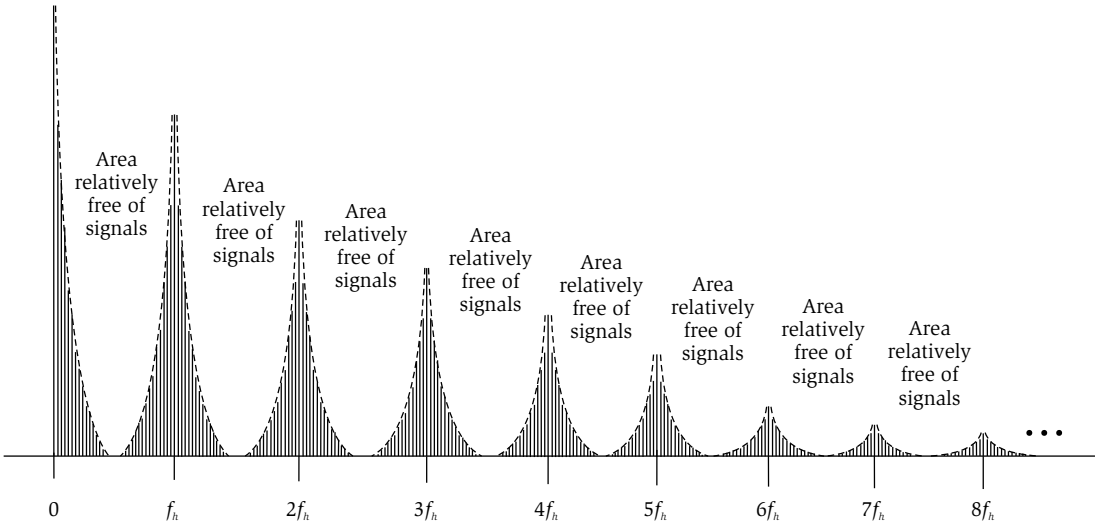


Figure 2.11 Video spectral energy clusters around harmonics of horizontal scan frequency, f_h, all the way up to the upper band edge.

For many years, audio in television was underappreciated. Only in the last decade has it received the attention it deserves as a fundamental part of the viewing experience. Perception experiments with human subjects at the Media Laboratory of the Massachusetts Institute of Technology have demonstrated that subjects shown two identical video presentations with differing accompanying sound qualities believed that the video with the better sound also had better video quality.

2.3 Modulation

Modulation is the process by which an information-bearing baseband signal is impressed upon a "carrier" signal. Baseband frequencies are those from 0 Hz to the maximum frequency content of the signals that come out of a television camera or a microphone. Only one baseband signal can be conveyed per transmission line, since a second signal, occupying the same frequency range, would interfere. Modulation allows multiple baseband signals to occupy different frequency bands in the cable through use of different carrier frequencies. The separately modulated signals are then added together to provide one composite signal for transmission. This process, called *frequency division multiplexing* (FDM), gives rise to the familiar concept of "channels." Each channel is a band of frequencies associated with one video and audio signal.

There are several ways of approaching the subject of amplitude modulation. A phenomenological approach describes the physical processes and follows the historical path by which amplitude modulation was invented and deployed. A mathematical approach starts from first principles and develops relatively simple equations describing amplitude modulation and detection and the special versions of each developed for NTSC television. This approach builds a firm foundation upon which to consider the addition of compatible color to the monochrome television system. Additionally, Chapter 8 explains the modulation process in terms of vector diagrams called *phasors*.

2.3.1 A Phenomenological Approach to Amplitude Modulation

A television camera produces signals that are proportional to the strength of the light in the image and includes synchronization pulses. Figure 2.8 shows the waveform of such a signal. This is an information-bearing baseband signal because the frequencies of this signal extend from zero to the maximum produced by the television camera. The zero-frequency component represents the average brightness of the image. The highest frequency conveys the finest detail of the picture.

A natural method of conveying the television camera signals would be to modulate the strength of a carrier wave in proportion to the amplitude of the original baseband signal, as shown in Figure 2.12.[9] The outline of the maximum

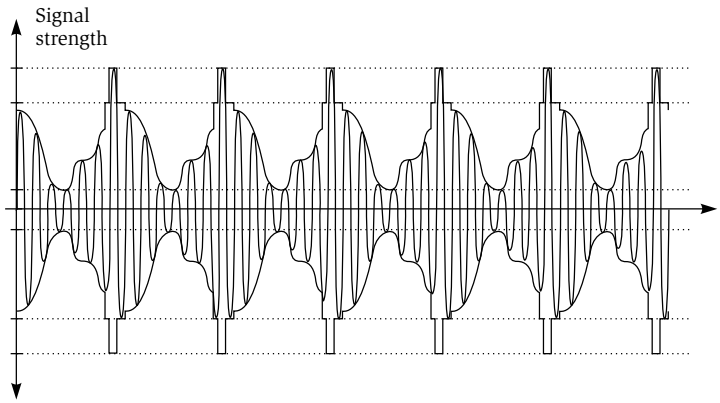


Figure 2.12 Television modulation. Carrier wave (shown at much lower frequency for convenience of illustration) modulated with video signal.

excursions of the carrier, termed the *envelope*, is seen to be a replica of the original baseband information-bearing signal.

At the receive site, the goal is to display the signal from one television camera. This can be done first with a tuned circuit that passes only the desired modulated carrier (actual receivers implement a more sophisticated version of this method based on the superheterodyne receiver principle), followed by a simple diode that will pass only the positive-going portion of the modulated signal, followed by a low-pass filter to remove the carrier pulses, as shown in Figure 2.13.

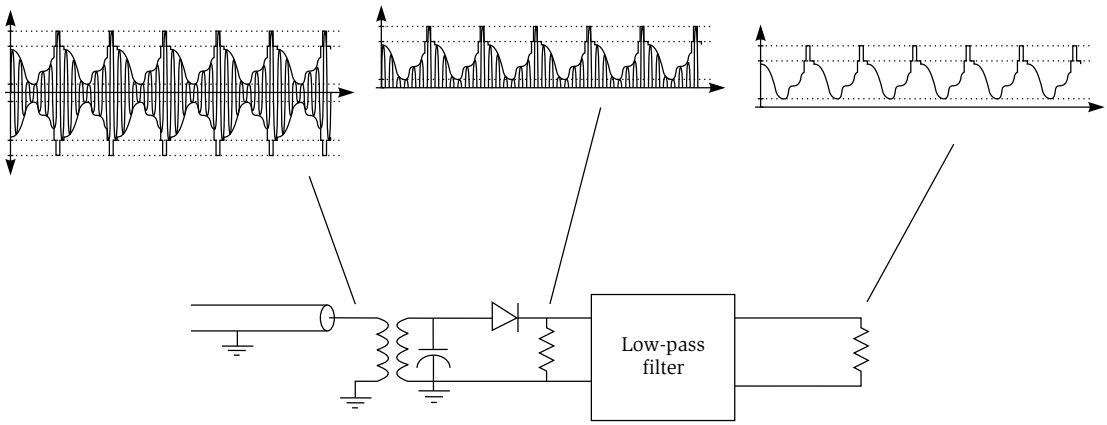


Figure 2.13 Envelope detection with a diode and low-pass filter.

2.3.2 A Mathematical Approach to Amplitude Modulation

Amplitude modulation is the process by which a high-frequency carrier wave is multiplied by a low-frequency information-bearing signal to generate a new signal, centered on the carrier frequency, which includes *sidebands* that carry the information. Through use of multiple carriers, each modulated by a different information-bearing signal, it is possible for multiple information streams to share a common transmission medium.

More formally, let us define the two signals as follows:

$$\begin{aligned} \text{Carrier} &= A\sin(\omega_c t) \\ \text{Information} &= (1 + B\sin(\omega_v t)) \end{aligned} \tag{2.4}$$

where:

$\text{Carrier} =$ carrier wave

$A =$ peak amplitude of the carrier wave

$\omega_c =$ angular frequency of the carrier wave $= 2\pi f_c$

$f_c =$ frequency of the carrier wave, in hertz

$\text{Information} =$ information-carrying signal (simplified to a simple sine wave to simplify the equations)

$B =$ peak amplitude of the information-carrying signals (must be less than 1)

$\omega_v =$ angular frequency of the video information

$t =$ time, in seconds

If we multiply these, we get

$$A[\sin(\omega_c t) + B\sin(\omega_c t)\sin(\omega_v t)] \tag{2.5}$$

That is, we get the original carrier plus the effect of multiplying the original carrier times the information-carrying signal. When we multiply two sine waves, however, we get two signals whose frequencies are equal to the sum and difference of the two original frequencies and whose amplitudes (voltage, in this case) are in accordance with the standard trigonometric identity:

$$\sin A \sin B = 0.5\cos(A - B) - 0.5\cos(A + B) \tag{2.6}$$

In other words, we get a sideband on each side of the carrier that is spaced from the carrier by the information-carrying frequency and whose amplitude will be $6 + 20$ log (B) dB less than the amplitude of the unmodulated carrier.

In the case where the information-carrying signal is a scanned television picture, it will have frequency components extending from zero hertz to an upper frequency limit that is related to the possible horizontal resolution of the displayed image. In the NTSC system, the information-carrying signal is limited to about 4.2 MHz of bandwidth. Thus, with the simple *double sideband* modulation just described, the modulated signal bandwidth would be 8.4 MHz wide just for video information, with the audio still to be fitted into the spectrum.

It is apparent, however, that double sideband modulation is wasteful of spectrum, because each sideband carries exactly the same information. An alternative is *single sideband* (SSB) modulation, whereby only one of the sidebands and no carrier is transmitted (saving both bandwidth and transmitter power). However, the detection process is more complex for SSB signals because a local carrier must be generated to mix with the incoming sideband in order to recover the original information-carrying signal. While this is simple today, it wasn't when the NTSC system was being developed and so was rejected.

The compromise reached was vestigial sideband (VSB) modulation, in which a normally modulated double sideband signal is passed through a filter that passes one sideband (the upper sideband of the transmitted signal), but only 0.75 MHz of the other sideband, as shown in Figure 2.14(a). As can be seen, the sound signal has been inserted 4.5 MHz above the visual carrier, leading to a total visual + audio signal that is only 6 MHz wide yet includes picture information extending up to between 4.0 and 4.2 MHz.

If this signal is demodulated using a standard amplitude detector, the result will be that modulating frequency components below 0.75 MHz will be recovered from both sidebands, while those falling above 1.25 MHz will be recovered from only a single sideband, thereby distorting the relative amplitudes of the detected information signal, as shown in Figure 2.14(b). In order to correct for this, the modulated signal must be passed though a compensating filter before detection. The shape of the required filter is shown in Figure 2.14(c).

2.3.3 Video Bandwidth

When the NTSC system was being designed, a tremendous number of human visual experiments were undertaken. Estimates were made on the likely course of technology. It is a tribute to the work done that the NTSC system has lasted for more than 50 years and is likely to be in use for at least another decade or two.

NTSC television is meant to be a "five times picture height" system. That is, the compromises made between video quality and economics were such that they would not be noticed if the picture were viewed from a distance of at least five times the picture height. Consideration of this and of the bandwidth available, as well as of the ways of occupying that bandwidth in a manner that would allow affordable receivers, led to the choice of 525 total (483 active) scan lines. If a synthetic video pattern is generated by electronically making alternate lines

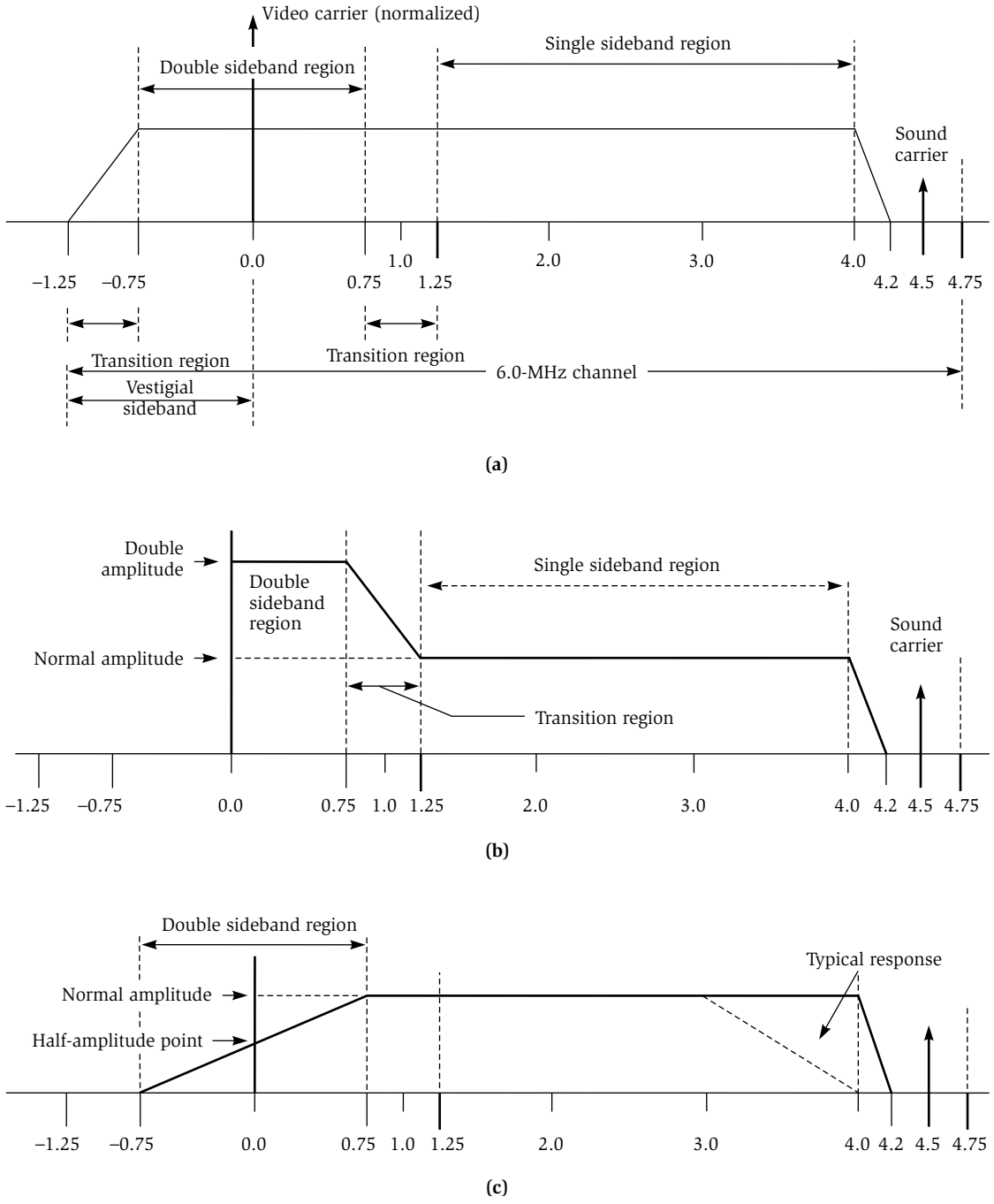


Figure 2.14 VSB reception. (a) Vestigial sideband spectrum normalized to 0.0 Hz as seen at RF (at IF, spectrum is reversed). (b) Output of a "flat response" television receiver as seen at RF (at IF, spectrum is reversed). (c) Idealized and typical TV receiver response curves as seen at RF (at IF, spectrum is reversed).

black and white, and if the whole raster is placed in the viewing area of the screen, then 483 picture elements can theoretically be resolved in the vertical direction. In the "real world," a large number of factors limit vertical resolution, including that the scanning system is a sampling system and that raster lines do not fall exactly on object boundaries. Rather than a clean edge, some of the light falls into two or more adjacent lines, blurring the image. Additionally, there is raster instability. Minor movements between scans add further blurring. The scanning spot in the pickup device and the scanning spot in the display device are not precise points but have a bell-shaped (Gaussian curve) distribution of effectiveness. All these complex factors yield less actual vertical resolution than the number of scan lines would indicate. The ratio between the number of scan lines and the actual vertical resolution varied from 0.5 to 0.85 in various experiments and is commonly accepted as 0.7. R. D. Kell is credited with much of the work on this phenomenon, so this ratio is called the *Kell factor*. That is, 483 active scan lines yield an effective vertical resolution of about 338 lines.

There is no single derivation of the bandwidth needed to carry video. The best that can be done is to give a plausibility argument for why something like 4.2 MHz is a reasonable number. Assuming a desire to have approximately the same horizontal resolution as vertical resolution and noting that the *aspect ratio* (the ratio of vertical height to horizontal width) of an NTSC picture is 3 by 4, we must multiply 338 by 4/3, yielding 450 horizontal lines of resolution. Dividing the aspect-ratio-adjusted number of horizontal picture elements by 2 yields the maximum number of cycles of video in the active horizontal video time. This yields 225 cycles. The 225 cycles must take place in 52.6 microseconds, yielding a frequency of 4.28 MHz.

The maximum number of picture elements possible in a video frame constructed electronically is 483 scan lines times 450 horizontal lines of resolution, or 217,350 pixels. European television systems that use the 50-pictures-per-second rate have more time per picture, allowing for 625 scan lines. To retain a balance between vertical and horizontal resolution, these systems have allocated more bandwidth per channel.

2.3.4 Frequency Modulation

The previous discussion of modulation considered only modifying the strength of a carrier signal to impress an information-bearing baseband signal upon it.[10] A cosine wave, however, has three parameters that can be varied to convey information. Equation (2.7) shows them:

$$s(t) = A \cdot \cos(2\pi ft + \theta) \tag{2.7}$$

Here A is the amplitude, f is the frequency, and θ is the phase angle of the cosine function. The symbol $s(t)$ is the signal as a function of time. As previously discussed, visual information is added to the carrier by varying its amplitude.

Sound information is modulated on a separate carrier by varying its frequency, in which case the signal can be expressed by

$$s(t) = A \cdot \cos(2\pi[f_1 + g(t)]t + \theta) \tag{2.8}$$

where the information-bearing signal is $g(t)$. Without loss of generality, the constant phase angle θ can be dropped, yielding

$$s(t) = A \cdot \cos(2\pi[f_1 + g(t)]t) \tag{2.9}$$

If the baseband information-bearing signal is a simple cosine wave of frequency f_2, the resulting expression is

$$s(t) = A \cdot \cos(2\pi[f_1 + \cos(2\pi f_2 t)]t) \tag{2.10}$$

The mathematics of cosines within cosines rapidly becomes very complex, involving advanced functions and concepts. There are excellent treatments of these details for anyone who is interested. The principle result for our purposes, however, is that the bandwidth of a frequency modulated signal can exceed that of an amplitude modulated signal given the same information-carrying signal. The signal, however, is more robust and can survive noise and other distortions better than amplitude modulation. Frequency modulation (FM) belongs to a class of signals called *spread spectrum* because it spreads the information signal over more bandwidth than it occupies in its baseband version. This technique trades off spectrum inefficiency for signal robustness.

The phenomenological concept motivating frequency modulation is the notion that noise, particularly static, was believed to be only an amplitude effect. If the amplitude of the signal did not carry information, the impact of noise added to the signal would be negligible. This simplistic model is mostly correct, and the main benefits of FM include much less susceptibility to noise.

The creators of NTSC considered it to be a video enhancement to radio. The intention was that, at the outer limits of reception, the audio signal would survive even if video failed,[11] and thus it was made extremely robust by using frequency modulation. In cable practice, we are able to take advantage of this by reducing the relative strength of the audio signal and thus reducing the overall signal loading on the cable system. In NTSC television, the sound signal is frequency modulated onto a carrier that is 4.5 MHz above the visual signal. Over-air broadcasters use separate transmitters for visual and aural signals and combine the signals before feeding a shared antenna.

2.4 Modern Analog Television

Modern analog television differs from the original system in only three respects: the addition of color, stereo sound, and ancillary signals. All this has been accomplished without making earlier products obsolete. If the very first commercial television receiver were operational and powered today, it would still receive and display television programming. "Compatibility," however, is an ambiguous term, and older receivers experience slightly reduced performance. These compromises stand the cost–benefit test and have provided the viewing public with a net positive benefit.

2.4.1 Color Television

Color is the most significant addition to NTSC.[12,13] A large number of trade-offs were necessary in the original NTSC system to make it practical and affordable, but even more were required to add color. The color system carefully matches its design to the human visual system to minimize the incremental information rate required.

The most important discovery is that the human visual system responds to color in two distinct ways: (1) as a response to the frequency (or wavelength) of the light, and (2) as a response to combinations of primary colors. If we pass sunlight through a glass prism, we decompose the light into individual wavelengths that we see as many separate colors. If this were the only way in which the eye perceives color, a color television system, or even color printing and photography, would be all but impossible. Fortunately, the eye, in response to various ratios of red, green, and blue,[14] perceives nearly all the possible colors.

The science of colorimetry is fascinating and complex and well beyond the scope of this brief chapter. Only a few basics can be covered here. In the color receiver, the picture tube is actually three picture tubes in one.[15] There are three electron guns and three sets of colored phosphor elements on the screen. A *shadow mask* in the tube is carefully aligned to allow only the electrons from a particular gun to strike the phosphors of its assigned color. The shadow mask is also called an *aperture mask* because of its array of holes, or apertures (see Figure 2.15). The three electron guns are tilted at an angle with respect to the axis of the picture tube. The three electron beams pass through the same holes in the shadow mask, which is located very close to the screen. Because they aim from different angles in the neck of the tube, the beams land on adjacent but different spots. The spots have phosphors that emit red, green, and blue light. The spots are small and close together, and the eye fuses them into a continuous color from normal viewing distances. The reader is invited to take a magnifying glass and look closely at the screen of a color receiver. Clusters of dots or clusters of small bars of the three different colors will be visible when the screen is illuminated.

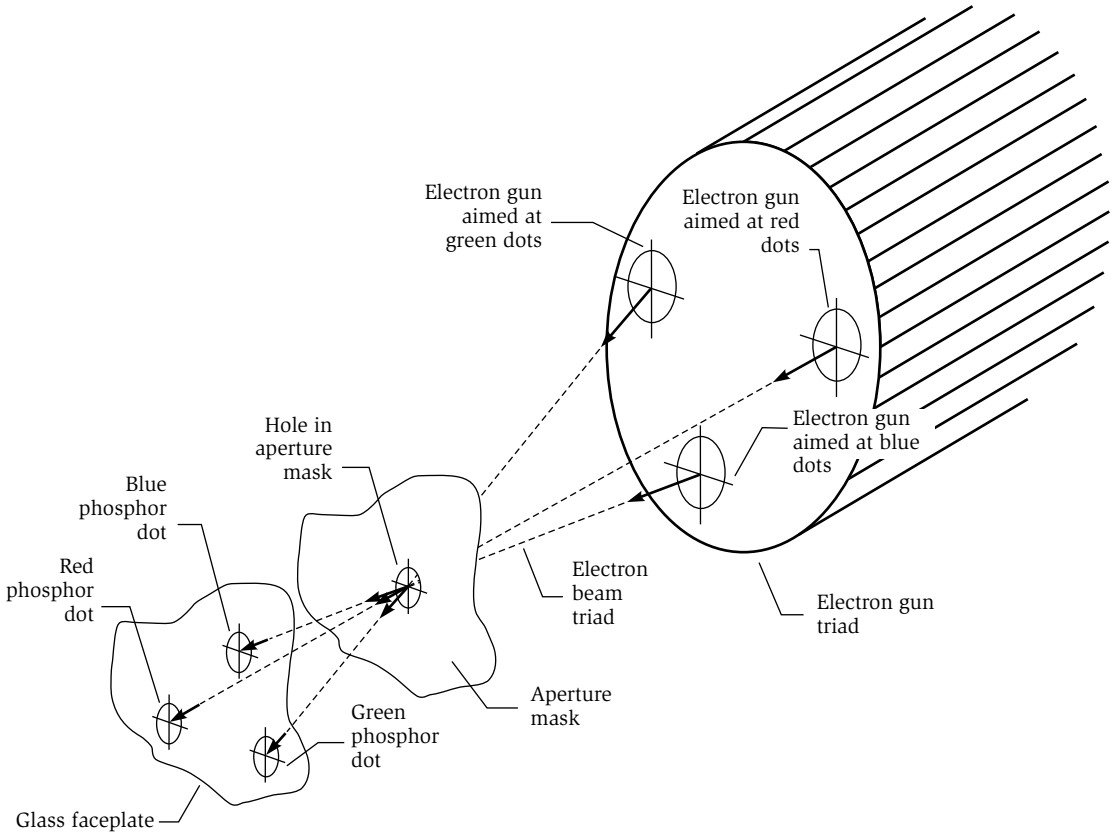


Figure 2.15 Color picture tube structure with colored phosphor dots, aperture (or shadow) mask, and cluster of three electron guns.

If a combination of three complete images in red, green, and blue were required to make a color picture, the required transmission bandwidth would be tripled. Even the improvement color brings might not be worth that drastic a loss of channel capacity. Fortunately, the eye is quite pleased with a "colored" television system and does not require a true "color" system. If we consider the Sunday color comics, we realize that we have a detailed black-and-white image with some color thrown on it. The eye and the imagination do the rest. In colored television, the same approach is taken. The detail is in the luminance (brightness) signal. The color signal contains much lower resolution and only approximately fills in the monochrome outlines. That this is the case can be seen when colored text is in the picture. For example, a yellow letter T on a blue background will have a yellow top bar because the image is large enough

horizontally to be represented by the frequencies involved. However, the vertical component of the T will usually not be yellow. To represent it in yellow would require frequency components beyond what is available.

To make the color signal compatible with older monochrome receivers, it was necessary to combine the signals in a manner that yielded the luminance signal plus two narrower bandwidth "color difference" signals. Since the eye is more sensitive to detail in certain colors than in others, it is possible to reduce the bandwidth of some of the colors more aggressively. The color television system manipulates the color signals using a set of linear matrix equations. This separates the signal into a component that contains flesh tones and near flesh tones and another component that has quite different colors. The principal color signal, termed the *I* signal, has a transmitted bandwidth of around 1.5 MHz. The secondary color signal, termed the *Q* signal, has a transmitted bandwidth of around 0.5 MHz. In a color receiver, linear networks can then separate the signals into the three primary colors, red, green, and blue. Simultaneously, a monochrome receiver will respond only to the luminance signal and ignore the color signals.

Although the NTSC system uses bandwidths of 1.5 MHz in the *I* channel and 0.5 MHz in the *Q* channel, practical color receivers actually implement a bandwidth of only about 0.5 MHz (or less) in each channel. This compares with several megahertz for the luminance signal. The main reason for the reduced bandwidth is the problem of "chroma noise." When the higher-frequency chroma signals are synchronously demodulated (a frequency conversion process), the noise added at those frequencies is also brought down in frequency. This causes the interference resulting from the noise to increase in size on the screen. Reducing the chroma bandwidth minimizes this unpleasant effect, but at a reduction in color resolution.

Even with only 1.5 MHz of *I* signal and 0.5 of *Q* signal, we still need to find a place to put 2.0 MHz of additional information. The principle of compatibility precludes simply adding more bandwidth. This is where the work of Mertz on the spectrum of scanned images provides an answer. Figure 2.11 showed that the signal energy is clustered around multiples of the horizontal scan frequency, with significant spectral gaps. Since this is a consequence of scanning and the color information likewise is a product of scanning, it too has spectral gaps. The trick then is to arrange the two spectra such that the energy clusters interleave. Figure 2.16 shows that if the color signal is modulated onto a carrier that is an odd harmonic of half of the line frequency, the color signal spectrum will interleave with the monochrome signal spectrum. The color signal occupies frequencies that are nearly empty in a monochrome signal.

This technique for hiding the chroma signal in the luma can be visualized in the time domain as well (See Figure 2.17). A cosine wave that is an "odd harmonic of half of the line frequency" will start with the opposite phase on adjacent lines. In the basic case, the first harmonic of half the line frequency will make only half of a cosine wave in the time of one horizontal scan. It will begin

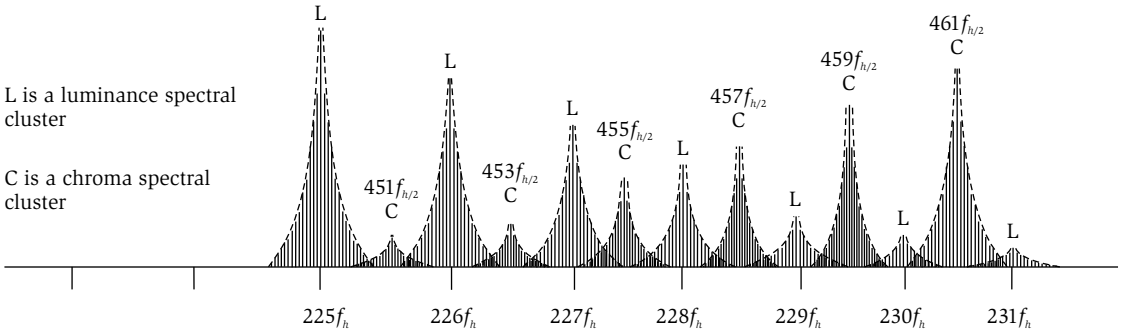


Figure 2.16 Luminance spectral energy clusters around harmonics of horizontal scan frequency f_h all the way up to the upper band edge. Chroma spectra energy components cluster around odd multiples of half horizontal scan frequency and are interleaved with luminance.

the next line with the opposite phase. So where the signal is brightest on one line, it will be dimmest on the adjacent line, and vice versa. These opposite brightnesses on adjacent lines will tend to average out and reduce visibility of the color carrier. The same is true of higher odd harmonics (even harmonics would cause the signal to have the same phase on adjacent lines, and the bright and dark spots would reinforce each other rather than tend to cancel). At the color subcarrier frequency, the bright and dark spots on the screen are very small and are positioned on adjacent lines relative to each other so as to average out the brightness they cause.

The two color signals, I and Q, are modulated in quadrature to each other onto the color subcarrier.[16] *Quadrature modulation* means that two carriers of the same frequency are used, one phase shifted by 90° from the other. The color signal can be expressed as

$$c(t) = I(t) \cdot \cos(2\pi f_c t) + Q(t) \cdot \sin(2\pi f_c t) \tag{2.11}$$

The sine function is 90° phase shifted from the cosine function. The required detection process is called *quadrature detection*. In the receiver, quadrature detection cleanly separates the two signals. This separation is based on the following simple trigonometric identities:

$$\cos(2\pi f_1 t) \cdot \cos(2\pi f_2 t) = \frac{1}{2}\cos(2\pi[f_1 - f_2]t) + \frac{1}{2}\cos(2\pi[f_1 + f_2]t) \tag{2.12}$$

$$\sin(2\pi f_1 t) \cdot \sin(2\pi f_2 t) = \frac{1}{2}\cos(2\pi[f_1 - f_2]t) - \frac{1}{2}\cos(2\pi[f_1 + f_2]t) \tag{2.13}$$

$$\sin(2\pi f_1 t) \cdot \cos(2\pi f_2 t) = \frac{1}{2}\sin(2\pi[f_1 - f_2]t) + \frac{1}{2}\sin(2\pi[f_1 + f_2]t) \tag{2.14}$$

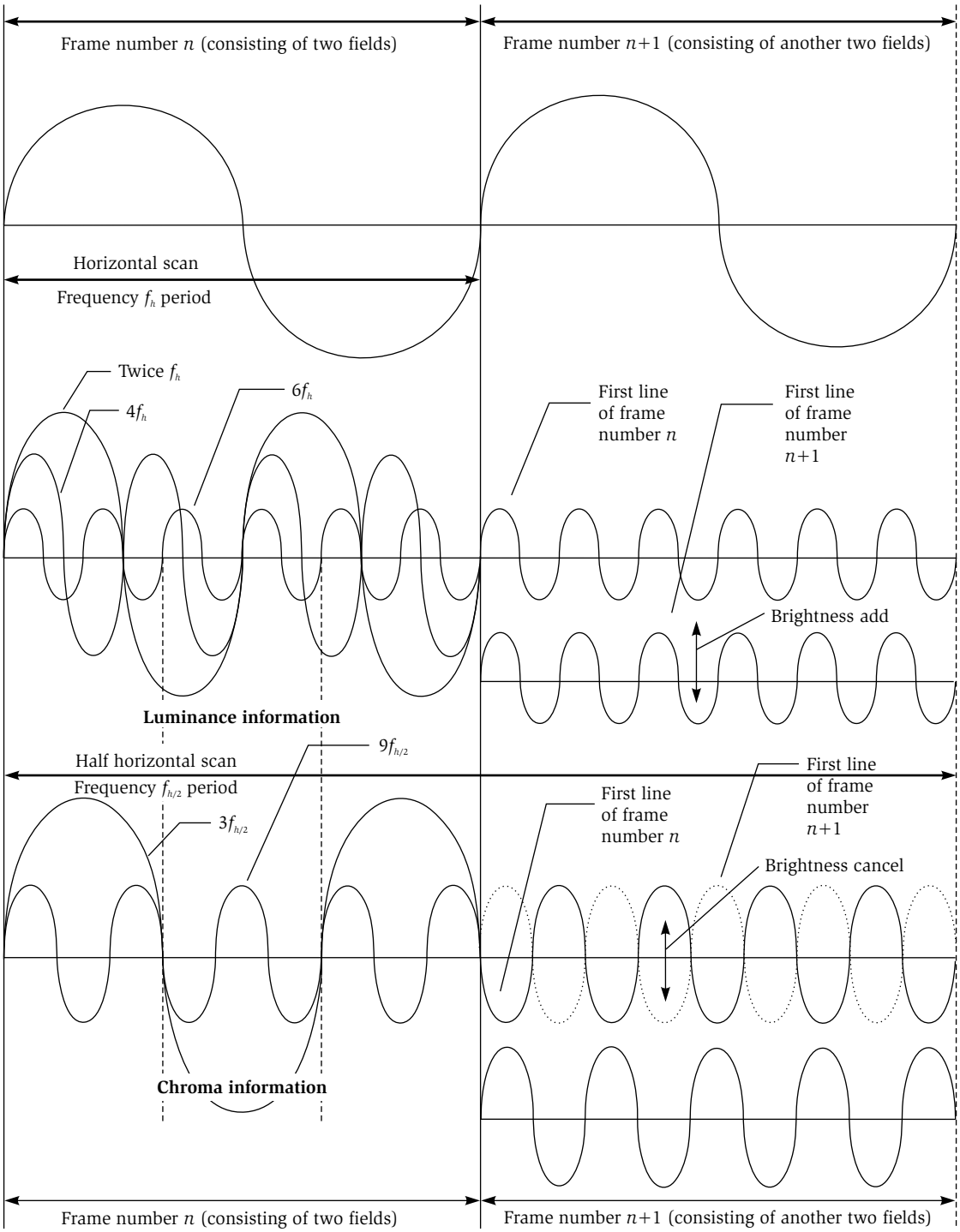


Figure 2.17 Time domain view of the frequency interleaving principle used to hide color information.

Let $f_1 = f_2 = f_c$, which is the carrier frequency for the color information. Then, from Equation (2.12),

$$\cos^2(2\pi f_c t) = \frac{1}{2}\{1 + \cos(2\pi[2f_c]t)\} \tag{2.15}$$

And from Equation (2.13),

$$\sin^2(2\pi f_c t) = \frac{1}{2}\{1 - \cos(2\pi[2f_c]t)\} \tag{2.16}$$

And from Equation (2.14),

$$\sin(2\pi f_c t) \cdot \cos(2\pi f_c t) = \frac{1}{2}\sin(2\pi[2f_c]t) \tag{2.17}$$

If we take the color information signal, $c(t)$, of Equation (2.11) and multiply it by $\cos(2\pi f_c t)$, a cosine at the color information carrier frequency, we get

$$c(t) \cdot \cos(2\pi f_c t) = I(t) \cdot \cos^2(2\pi f_c t) + Q(t) \cdot \sin(2\pi f_c t) \cdot \cos(2\pi f_c t) \tag{2.18}$$

Applying Equation (2.15) to the $I(t)$ term and Equation (2.17) to the $Q(t)$ term, we get

$$c(t) \cdot \cos(2\pi f_c t) = \frac{1}{2}I(t) \cdot \{1 + \cos(2\pi[2f_c]t)\} + \frac{1}{2}Q(t) \cdot \sin(2\pi[2f_c]t) \tag{2.19}$$

$$c(t) \cdot \cos(2\pi f_c t) = \frac{1}{2}I(t) + \frac{1}{2}\{I(t) \cdot \cos(2\pi[2f_c]t) + Q(t) \cdot \sin(2\pi[2f_c]t)\} \tag{2.20}$$

The first term on the right-hand side of Equation (2.20) is the $I(t)$ signal at baseband frequencies. The bracketed terms are modulated onto carriers at twice the chroma carrier frequency, f_c, and can easily be removed with a low-pass filter. Thus, the $I(t)$ signal has been completely separated from the $Q(t)$ signal through quadrature detection.

If we take the color information signal, $c(t)$, of Equation (2.11) and multiply it by $\sin(2\pi f_c t)$, a sine at the color information carrier frequency, we get

$$c(t) \cdot \sin(2\pi f_c t) = I(t) \cdot \sin(2\pi f_c t) \cdot \cos(2\pi f_c t) + Q(t) \cdot \sin^2(2\pi f_c t) \tag{2.21}$$

Applying Equation (2.17) to the $I(t)$ term and Equation (2.16) to the $Q(t)$ term, we get

$$c(t) \cdot \sin(2\pi f_c t) = \frac{1}{2}I(t) \cdot \sin(2\pi[2f_c]t) + \frac{1}{2}Q(t)\{1 - \cos(2\pi[2f_c]t)\} \tag{2.22}$$

$$c(t) \cdot \sin(2\pi f_c t) = \frac{1}{2}Q(t) + \frac{1}{2}\{I(t) \cdot \sin(2\pi[2f_c]t) - Q(t) \cdot \cos(2\pi[2f_c]t)\} \tag{2.23}$$

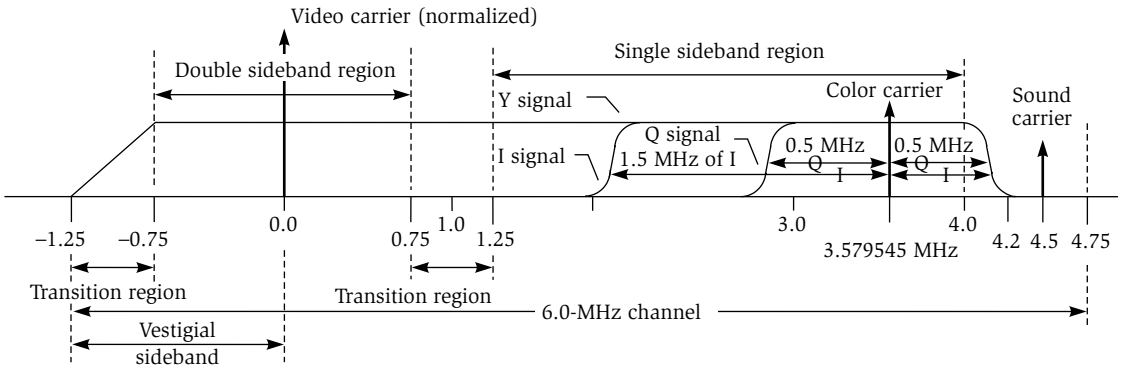


Figure 2.18 Location and bandwidths of the luminance and chroma signals.

As with the $I(t)$ signal, the $Q(t)$ signal has been completely separated through quadrature detection.

The I signal is the *in-phase* color signal and is amplitude modulated onto a color subcarrier that is located at 3.579545 MHz, the 455th odd multiple of half the line-scanning frequency. The Q signal is the *quadrature* color signal. It is amplitude modulated on a carrier that is derived by a 90° phase shifting of the color subcarrier (see Figure 2.18). Note that the Q signal is a double sideband signal, whereas the I signal has a 0.5-MHz double sideband component and an additional 1.0 MHz of single sideband modulation. However, as already mentioned, practical consumer receivers limit bandwidth in both channels to around 0.5 MHz.

It will be recalled from the study of geometry that there are two common forms of coordinate systems, rectangular and polar. In the rectangular system, a location is defined by the distance in the horizontal direction and the distance in the vertical direction. In the polar system, a location is defined by a distance along an angle. This is like giving driving directions of 3 miles east and 4 miles north. If the traveler has an airplane, he can simply be told to fly 5 miles on the heading of 30°. The diagram of Figure 2.19 applies these principles to the quadrature color equation (2.11). The color signal $c(t)$ can be given either as the sum of the $I(t)$ signal and the $Q(t)$ signal or as the resulting magnitude $c(t)$, which is the square root of the sum of the squares of $I(t)$ and $Q(t)$ at an angle given by the inverse tangent of the ratio of $Q(t)$ to $I(t)$. In this "polar" way of interpreting the color signal, the color, or hue, is defined by the angle, and the intensity of the color (termed its *saturation*) is described by the length $c(t)$. Figure 2.20 shows the location of various colors on a polar graph.

The changes made to accommodate color required the horizontal and vertical scan rates to be slightly modified. This is because the difference between the color subcarrier frequency and the aural subcarrier frequency is a frequency well within the video bandwidth. Any nonlinearity in that signal path will produce a

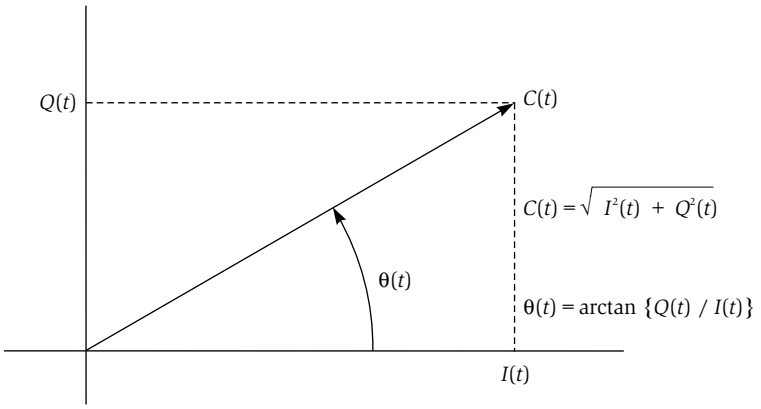


Figure 2.19 Rectangular versus polar representation.

	C	θ
(B–Y)	1	0°
Q	1	33°
Magenta	82	61°
(R–Y)	1	90°
Red	88	103°
Yellow	62	167°
Burst	1	180°
Green	82	241°
Cyan	88	283°
–I	1	303°
Blue	62	347°

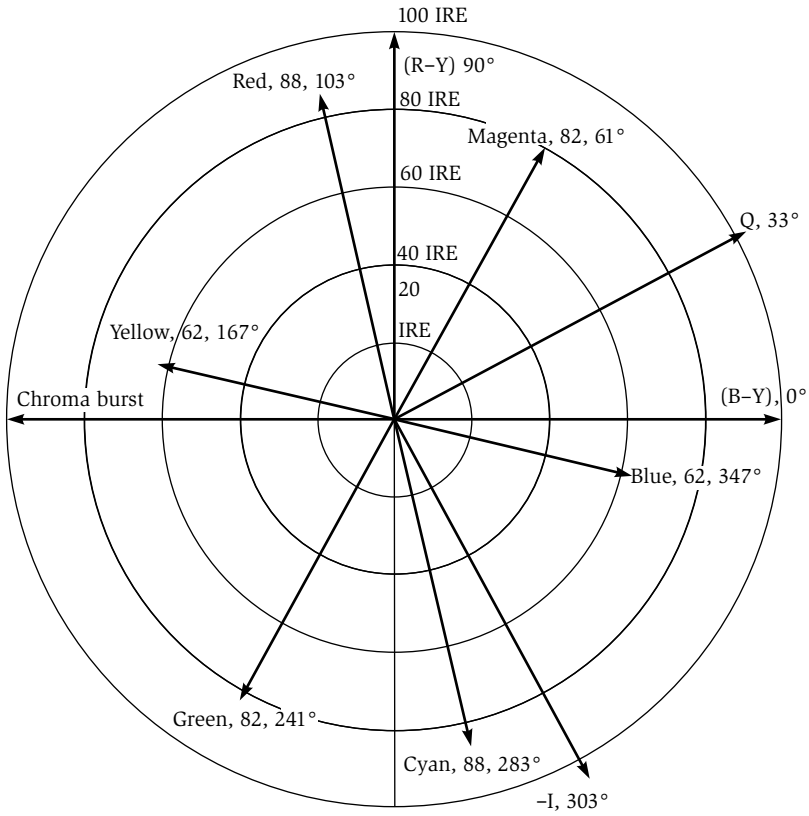


Figure 2.20 Polar representation of television colors.

beat product that is in the video passband and can produce an annoying pattern on the screen. This difficulty has been termed the 920-beat problem because the difference between the frequencies selected for the chroma subcarrier and the sound subcarrier is 920 kHz. Two techniques are used to minimize this problem. The existing 4.5-MHz trap strongly attenuates the aural carrier frequency before the video path. However, it was thought that this measure may not be sufficient in existing monochrome receivers that were designed and produced before the color standard was established. The remaining 920-kHz beat was made less visible by causing it to be frequency interleaved (just like the chroma signals) with the luminance signal. This required that the 920-kHz beat be made an odd multiple of half the scan frequency. To accomplish this (since the chroma subcarrier is already an odd multiple of half scan line frequency), the sound carrier must be any multiple of the horizontal scan frequency (that is, an *even* multiple of half scan line frequency). The 286th multiple of the monochrome horizontal scan frequency (15.75 kHz) is 4.5045 MHz. Using these numbers would require changing the aural subcarrier by 4.5 kHz. Some of the then-existing monochrome television receiver's audio circuits might have failed to perform with this much of a shift in frequency, jeopardizing the compatibility mandate. Also, the effectiveness of the aural frequency trap would be reduced. The alternative was to change the horizontal frequency so that it was the 286th submultiple of the original 4.5-MHz aural subcarrier. This yields a horizontal frequency of 15,734.26 Hz. Since there are 525 scan lines in a frame and two fields in a frame, the field rate will be the horizontal rate divided by 525 and multiplied by 2, yielding 59.94 Hz. The chroma subcarrier is the 455th multiple of half of this line rate, or 3.579545 MHz.

Synchronous detectors in receivers are used to completely separate the two chroma signals. Then, using circuits that perform the functions listed in Equations (2.24) through (2.26), the luminance signal E_Y, the in-phase color signal E_I, and the quadrature signal E_Q are matrixed to yield the blue signal E_B, the red signal E_R, and the green signal E_G that drive the tricolor picture tube.

$$E_B = E_Y + 1.70E_Q - 1.10E_I \tag{2.24}$$

$$E_R = E_Y + 0.62E_Q - 0.96E_I \tag{2.25}$$

$$E_G = E_Y - 0.65E_Q - 0.28E_I \tag{2.26}$$

Synchronization of the color oscillator used in the synchronous detection process is facilitated by the addition of a burst of unmodulated color subcarrier on each horizontal blanking pulse after the horizontal sync pulse (see Figure 2.21).

These techniques allow color television to be compatible with monochrome television. This compatibility is not complete because monochrome receivers built before the introduction of color had video bandwidths of up to 4.2 MHz, providing very sharp black-and-white pictures. When color signals were introduced, these receivers suffered from "dot crawl." The color signal was not

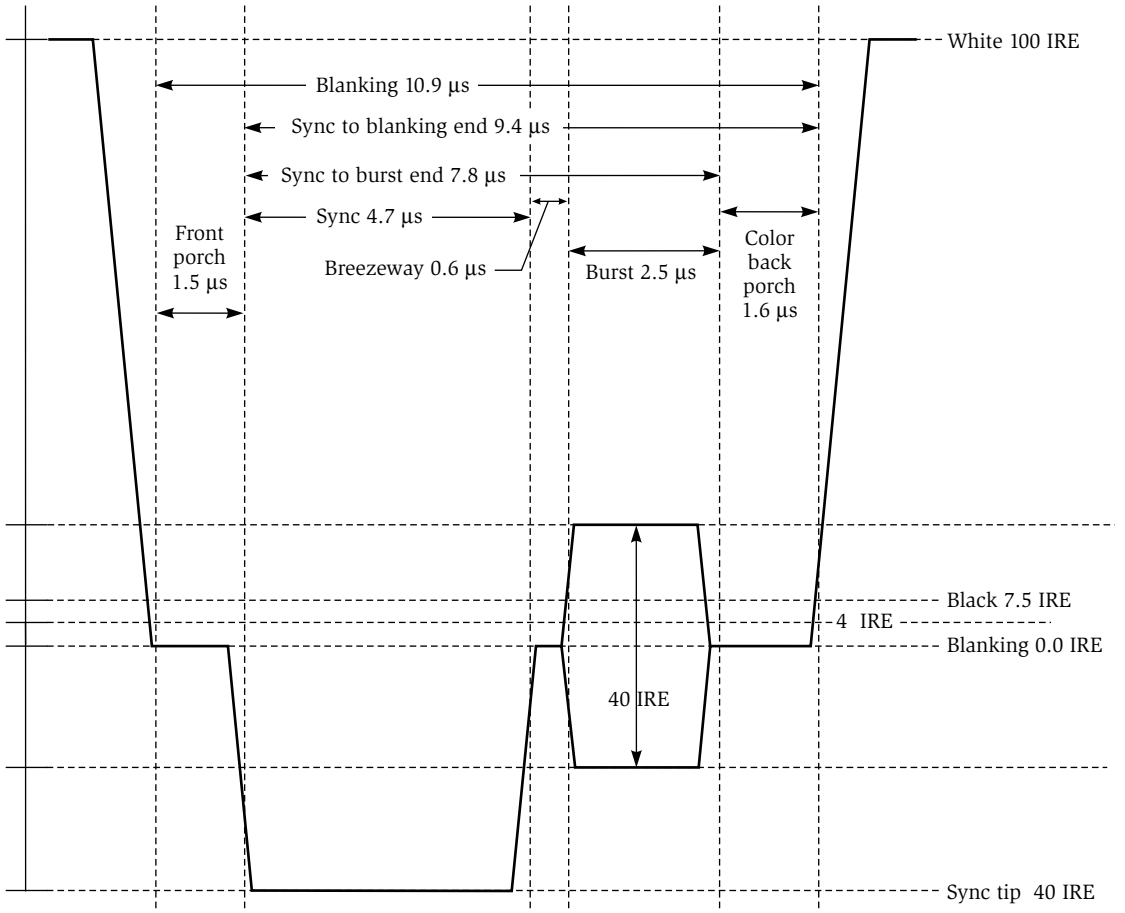


Figure 2.21 Color burst relationship to sync pulse.

adequately rejected by the older receivers and appeared as a moving pattern of annoying dots. This was overcome in later monochrome receivers by introducing a notch in the frequency response that eliminated much of the color signal but reduced sharpness. Alternatively, the video bandwidth of monochrome receivers was rolled off, with much the same effect.

Adequate compatibility was critical to the rational introduction of color television into a market already populated with monochrome television receivers. Consumers who purchased color receivers derived more benefits from the same signals that consumers with monochrome receivers continued to enjoy. Those who could not afford a color receiver could buy a new monochrome receiver and still have access to television. No one was disenfranchised by the technological advance to color.

2.4.2 Stereo Sound

The next improvement to NTSC came in the area of sound.[17] Television sound is frequency modulated on a separate carrier that is a fixed 4.5 MHz above the visual carrier. When stereo sound was added to the television system, the requirement of compatibility was again enforced. This avoided the chaos that would have resulted from making obsolete the existing sound system. Just as with compatible color, there were compromises to monaural receivers when stereo sound was added. But the net benefit to consumers was considered positive. The marketplace gave its approval to both compatible color and compatible stereo sound.

Stereo sound is implemented by first creating a spectrum that includes the sum of the left and right sound channels at baseband. As in FM radio broadcasting and monophonic television practice, the sum signal is preemphasized using a 75-microsecond preemphasis curve. (See the discussion in Chapter 8 under modulators.) The difference between the left and right channels is double sideband suppressed carrier modulated onto a carrier at twice the horizontal scan frequency ($2 \times 15{,}734 = 31{,}468$ Hz). In order to improve the signal-to-noise ratio, this difference signal is *companded*; that is, the signal is compressed in amplitude at the transmission point and expanded at the receiver. The compression is frequency selective and is based on an algorithm originally proposed by the dBx Corporation, whose name it bears. This is in contrast to the FM radio stereo system, which is not companded but employs the same deviation of the sound carrier, by both the sum and difference carriers. Both changes to the television system were brought about by the need for improved signal-to-noise ratio. A limited-bandwidth (10-kHz) monaural second audio program (SAP) channel is frequency modulated onto a carrier at five times the horizontal scan frequency. The SAP channel is intended for a second language or other such purposes. A very narrow-bandwidth (3.4-kHz) "professional channel" is frequency modulated onto a carrier at six and one-half times the horizontal scan frequency. It is used for television plant intercommunications. This entire complex spectrum is then frequency modulated onto the 4.5-MHz carrier (see Figure 2.22). The relationship between the video and aural carriers is tightly controlled, since nearly all television receivers depend on this relationship. The visual carrier is used as the local oscillator to bring the aural spectrum down to baseband. This technique is called the *intercarrier-sound method* of TV receiver design. Since the final aural modulation process is that of frequency modulation, the TV receiver uses a "limiter" circuit to strip off any amplitude modulation that may have become impressed upon the aural signal. Typical sources of this amplitude modulation are noise, particularly impulse noise, and cross modulation in nonlinear elements in the signal path. The TV receiver's aural system then becomes relatively insensitive to amplitude modulation on the aural carrier.

A variation of the previously described aural recovery technique is the *direct-sound method*. It has also been called the split-sound method. In this approach, the aural component is extracted from the composite IF signal prior

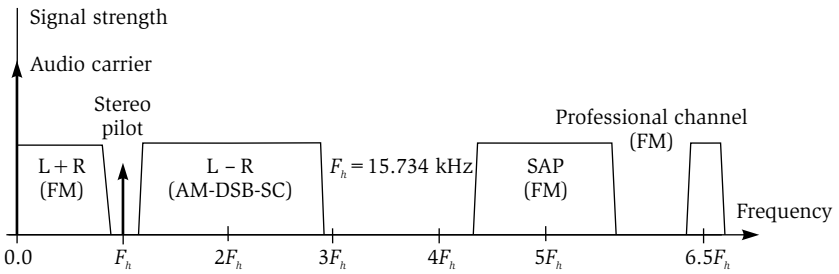


Figure 2.22 Stereo audio baseband spectrum.

to video detection. This signal is subjected to limiting to remove incidental amplitude modulation components and presented to an FM detector.

When direct-sound techniques are applied, the aural signal is not dependent upon any properties of the visual carrier, as it is in the intercarrier-sound approach, but is now affected by the accumulated instability of processing oscillators (heterodyning) along the signal chain. These sources of instability include the transmitter, cable heterodyne processors, and local oscillators of intervening frequency translation devices, such as cable subscriber terminal devices and the tuner of the TV itself.

There are trade-offs in making a choice between intercarrier-sound and direct-sound designs. Some visual modulators suffer from nontrivial amounts of incidental phase modulation of the visual carrier. The intercarrier principle causes these visual carrier instabilities to be transferred to the aural carrier, and this can introduce sound distortions. A direct-sound design would not be subject to this problem. Essentially all consumer TVs and VCRs use the intercarrier principle.

2.4.3 Ancillary Signals

It was determined that more uses can be made of the electron blanking time. One important application is to use the VBI to carry analog test signals. The analog test signals can be used to measure the transmission characteristics from the signal source to intermediate points along its path to the final point of use. Another use for the VBI is to carry analog signals that represent digital data. The data signals can be of two or more levels that are resolved into data bits by appropriate circuits. Since the "digital" signals are of just a few discrete levels, the data detection circuits can discriminate against significant amounts of noise, distortion, and interference. This makes these data signals more robust than the analog video signal itself for most forms of interference.

The first U.S. attempt to use the VBI for ancillary data purposes was in 1970, when the National Bureau of Standards (NBS — now the National Institute of

Standards and Technology, NIST) proposed to use it for the distribution of precise time information nationwide. The American Broadcasting Company (ABC) television network was a partner in that effort. Even though this initiative did not result in service, ABC recommended a captioning service for the hearing impaired.

Eventually the FCC reserved line 21 of field one of the television signal for the transmission of *closed captioning* (CC) in the United States in 1976. In 1990, Congress passed the Television Decoder Circuitry Act, mandating that after July 1, 1993, new television receivers of 13-inch diagonal display measure or greater must include caption decoding circuits. Approximately 20 million television receivers per year are covered by this requirement. In 1992, NCI worked with the FCC and the Electronic Industries Association (EIA) to develop captioning technical standards. The 1996 Telecommunications Act required the FCC to institute rules requiring closed captioning on video programming but allowing exemptions for programming that would suffer an "undue burden."

The CC system supplies data to appropriate digital and analog circuits that place carefully timed text on the television screen, making it possible for the hearing impaired to read a description of the conversation taking place and have indications of other relevant sounds. Similarly, those who cannot understand the spoken words may have text in their language, allowing them to follow the program. The CC system uses very low-speed data in order to minimize the impact of transmission path problems such as reflections and interfering signals. The data rate for the CC system is 503.5 kb/s of binary (two-level) data. This allows only two 8-bit characters to be transmitted per VBI line. If only field one is used, there will be 30 of these lines per second. This yields 480 bps, or 3,600 characters per minute. If the average word is 5 characters long with a space, then 600 words can be conveyed per minute. The rest of the line is occupied with both a burst of seven sine wave cycles of 503.5-kHz clock run-in and a unique "start bits" pattern placed at the beginning of the line. These signals synchronize the detector circuitry. Since only line 21 is protected for captioning by FCC rule, the rate of transmission is slow, but it is perfectly adequate for the purpose. The on-screen display consists of a maximum of 15 rows of 32 characters each. The captions usually appear only on rows 1 through 4 and rows 12 through 15. The middle rows are usually transparent to show the action. A text mode provides scrolling text. Further details can be found as part of EIA standard number EIA-608.

The closed captioning signal carries four components. There are two captioning channels and two text channels. The first captioning channel is synchronized to the video programming, causing the words to carefully match the video. The second captioning channel is not synchronized.

The FCC expanded the captioning standard EIA-608 to allow use of line 21 of field two. This adds two more captioning channels and two more text channels. A fifth channel has been added to carry extended data services (EDS). EDS can carry a wide variety of additional information, including precise time information to set clocks in consumer products. Finally, the original intent of the NBS will be realized.

The channel's name and call letters are included along with current program information, such as title, length, rating, elapsed time, types of audio services and captioning services, and intended aspect ratio. Also included is the data for the "V-chip" (violent programming advisory), which is intended to control children's access to programming that parents deem objectionable. Public service announcements such as weather and emergency advisories will be transmitted. Cable system channel layout information will be provided, causing the channel number indicator to use the more familiar channel identification number rather than the number associated with the frequency utilized. This facility can bring the same "channel mapping" benefits subscribers have enjoyed in their cable set-top terminals to consumer electronics products.

A subsequent VBI data transmission system, called *teletext*,[18] was invented to provide ancillary services to television users. The teletext system can display up to 24 rows of 40 characters on the television screen (but a specification of 20 rows was selected for use in the United States). Teletext quickly evolved into a transmission system for more complex data, including the downloading of software to computers. At the time teletext was introduced, electronic components were still relatively expensive, which severely limited the affordable capability of the teletext systems.

Teletext is a more aggressive form of data transmission than CC. It has been successful in Europe but has failed to enjoy commercialization in the United States. Teletext originated in Great Britain, with experimental transmission commencing in 1972. The British Broadcasting Corporation (BBC) branded their service Ceefax, and the Independent Broadcast Authority (IBA) called their service Oracle. France developed a packet-based teletext system called Antiope, based on a transmission system called Didon. Later, Canada developed another system, called Telidon, which featured higher-resolution graphics. The Japanese system, called Captain, featured "photographic coding" to accommodate Chinese Kanji characters and the Japanese Kana set.

There are a number of reasons for the difficulties in the United States. Principal among these was the failure to find a strategy that made money. Without this, the system could not be supported. Additional difficulties included the high cost of memory at the time of implementation. Even though a teletext page requires only about a kilobyte of storage, that small amount of memory was expensive then. Further problems centered on the quality of the graphics. The less expensive World System Teletext (WST), based on the British approach, had crude "Lego-style" graphics in its basic form. The other contender, the North America Presentation Layer Protocol System (NAPLPS), used a higher-resolution graphics system that painted itself on the screen painfully slowly, trying the patience of the user. Still another complication was the FCC's 1983 decision to allow two standards and hope that the marketplace would resolve the confusion. One of the systems is WST, and the other is the NAPLPS evolution of Antiope, Telidon, and efforts by AT&T. A final problem was reliability of data reception. In a test in the San Francisco Bay Area, only about 25% of installations of the

NAPLPS system were trouble free. The remainder suffered from various degrees of multipath impairment. The more robust WST system was not tested in that environment.

Both U.S. teletext systems have a data rate of 5.727272 Mbps, which is 364 times the horizontal rate and 8/5 of the color subcarrier. The data signal has a nonreturn to zero (NRZ) binary format. The WST data line consists of eight cycles of clock run-in (16 bits) followed by a unique 8-bit "framing code" followed by 16 bits of control codes and a payload of 32 8-bit display words. The displayed row has 40 characters, 8 more than can be conveyed in one VBI line of data. This is accommodated by transmitting 4 rows of 32 characters on 4 VBI lines and a supplementary VBI line with the 4 groups of 8 characters needed to finish the previous 4 display rows. This process is called *gearing*, because of the ratios between the display and the transmitted VBI data. Since the page format is 40 characters by 20 rows, with an additional "header row," 21 field lines are required plus 5 field lines of supplementary 8-character groups, for a total of 26 field lines per full page of WST teletext. The payload of 296 bits per line allocated means that if one VBI line in each field is allocated, a data rate of $296 \times 2 \times 30 = 17{,}760$ bps is obtained. Ten lines of VBI are possible (line 21 is reserved for captions, and the first nine lines form the vertical synchronization pulses), yielding a maximum of 177.6 kb/s for full VBI utilization.

The WST system maps the data location in the VBI line to memory locations and to screen locations, and always puts data in the same memory place. This allows for a very simple error-protection scheme. Since the instructions in the header are Hamming-code protected, a measure of the quality of the received signal is obtained. If the signal is of low quality, it is not loaded into memory. Only good-quality data is stored. As a result, good data can be accumulated from repetitions of the page until a good page of data is built up. It is also possible to use a "voting" approach to obtain very robust transmission.

The fundamental difference between the WST and the evolving set of Antiope, Telidon, and NAPLPS systems is that the latter systems all used a packet structure. They have been characterized as asynchronous because there is no mapping between the transmission scheme and memory and screen locations.

The Public Broadcasting System (PBS) has developed a packetized data delivery system based on teletext called the PBS National Datacast Network. The standard teletext data rate of 5.72 Mbps is used, yielding 9,600 baud per VBI line allocated per field. There are a wide variety of commercial applications for this signal. Currently, the StarSight Electronic Program Guide (EPG) signal is distributed via PBS.

A system by the WavePhore Corporation utilizes a teletext-like system in lines 10 through 20 in each field, for a data speed of up to 150 kb/s. WavePhore added substantial error-detection and -protection bits to its structure to protect against multipath and other transmission problems.

Underutilized portions of the NTSC spectrum can be employed to "hide" data. In many cases, the process of hiding the data is incomplete, and results in

artifacts under certain conditions. In other cases, the preparation of the NTSC signal to better hide the data reduces video quality. So the challenge is to both hide the data and not impair video quality while retaining signal robustness and the potential for an economic implementation.

The National Data Broadcasting Committee (NDBC) was formed in 1993 to establish a single standard for data transmission in video. On June 28, 1996, the FCC approved digital data transmission in the video portion of broadcast television transmission in its Report & Order (R&O) "Digital Data Transmission Within the Video Portion of Television Broadcast Station Transmissions," in MM docket No. 95-42. It allows four formats: two of the formats, by Yes! Entertainment Corporation and A. C. Nielsen Co., place low-data-rate signals in the overscan region of the picture; the other two systems, Digideck and WavePhore, embed the digital signal into the video signal.

The WavePhore system begins by reducing video luminance and chrominance bandwidths. The "luminance" is reduced from its theoretical value of 4.2 MHz to a lower value of 3.9 MHz, and the upper sideband of the color signal is reduced by about 300 kHz. It is then possible to insert a data signal in this region at a carrier frequency of approximately 4.197 MHz above the video carrier and a strength approximately 20 dB above the noise floor of the video system. The data is synchronous with the video carrier and thus with the horizontal line frequency. As an odd multiple of one-quarter of the horizontal scan frequency, it interleaves between the luminance and chrominance bundles of spectral energy. Data is not sent during the vertical and horizontal blanking intervals. Thirty bits of data are sent per video line. There are 240 available lines per field (not counting the VBI, during which the signal is blanked). This yields a raw data rate of 431.6 kb/s. After error-correction coding, the raw data rate is reduced to approximately one-fourth of a telephone T1 rate, or 384 kb/s, and referred to by WavePhore as their system TVT1/4.

WavePhore shuffles the data before applying biphase modulation and filtering out the lower sideband. Shuffling the data reduces its visibility in the video. An adaptive equalizer is used in the receiver. A major advantage of the WavePhore approach is that once inserted into the video, the data can be conveyed through the video path without giving it further attention. The WavePhore VBI system and the WavePhore subvideo system can be combined to provide more than 500 kb/s.

There is some degradation of pictures using this system. However, the FCC is willing to let the broadcaster determine what its individual marketplace values are and to respond to that decision.

The Digideck system adds a differential quadrature phase shift keyed (DQPSK) signal carrying about 500 kb/s placed 1 MHz below the video carrier. In this regard, it is similar to the European NICAM system for adding digital audio to analog television broadcasts (see Chapter 7). This places the new carrier in the vestigial side band region of the signal. To accommodate this, the lower VSB slope is increased. Rather than starting at the traditional 750 kHz below

picture carrier, in the Digideck system it starts at 500 kHz and drops more rapidly. The carrier is about 36 dB below peak power and has a raw capacity of 700 kb/s. Forward error correction and other overhead burdens reduce the data capacity to around 500 kb/s. Digideck calls the new carrier the D-Channel. The data signal is clocked synchronously to the television signal for ease of recovery and for better hiding in the video. The Digideck receiver also depends on an adaptive equalizer. A consequence of the D-Channel is that it must be inserted at the transmitter site and be brought there by an alternative path. Like the WavePhore system, Digideck introduces some artifacts. A marketplace approach will allow the broadcaster to determine acceptability.

The Yes! Entertainment Corporation's system introduces a pulse in the video between 9.1 and 10.36 microseconds after the start of the horizontal synchronization pulse. The data rate is very low, about 14 kb/s. Its application is to deliver audio to a talking toy teddy bear! A. C. Nielsen uses line 22 of one field of the video for transmitting a program source identification. This ID is used to measure the viewing population for statistical purposes.

EnCamera Sciences Corporation has presented papers on a system that compatibly embeds up to 4.5 Mb/s in an analog 6-MHz NTSC television signal. Up to 3.0 Mb/s is modulated using multilevel signaling onto a carrier that is in quadrature with the visual carrier. This approach would normally result in visible artifacts in the analog video. However, techniques have been developed that avoid these problems. An additional 1.5 Mb/s is amplitude modulated onto the aural carrier using multilevel signaling and specialized filters.

2.5 Signal Quality and Customer Expectations

The human response to television pictures is a complex phenomenon. This makes establishing performance specifications difficult. Yet these specifications are important for setting design goals, purchasing objectives, contractual obligations, and regulations.

Psychophysical testing assesses subjective responses of a large number of subjects to television pictures and assigns figures of merit versus objectively measured parameters such as random noise, phase noise, microreflections, composite triple beat, and chroma/luma group delay.

2.5.1 CableLabs Updates Early Testing

Early testing was done by the Television Allocations Study Organization (TASO) in 1958.[19] Further testing was done in 1983 by the network labs of the Columbia Broadcasting System (CBS), the National Aeronautics and Space Administration (NASA), and the Canadian Research Council (CRC).[20] Most recently, in 1991, work sponsored by CableLabs added to the base of knowledge concerning signal quality and customer expectations.[21]

The most significant conclusion is that customer expectations are rising. The TASO study proclaimed that in 1958 consumers found a weighted signal-to-noise ratio of about 30 dB to be "somewhat objectionable." In 1983, subjects' expectations were for a much better picture. About 40-dB signal-to-noise ratio was considered to be "somewhat objectionable." The CableLabs study found 45–46 dB to be considered "slightly annoying." This is a very substantial increase in demand for picture quality.

The CableLabs tests were conducted under carefully controlled circumstances and used both expert and nonexpert viewers. The expert viewers served primarily to set the endpoints of the scales for the different parameters. The selected visual material was presented on a custom-made analog video disc and on D2 tape.

Well-established procedures defined by the International Consultative Committee for Radio (CCIR) were employed.[22,23] The subjective scale ranged from "imperceptible," to "perceptible, but not annoying," "slightly annoying," "annoying," and "very annoying." The double-stimulus pair-comparison method used allows subjects to compare pictures to reference images. This eliminates the need to remember what images looked like at other points in time.

Subject material was selected for sensitivity to the subjective impairments under study. Sources for the material included the Society of Motion Picture and Television Engineers (SMPTE), the European Broadcast Union (EBU), and Kodak Corporation. All these entities have long histories in the analysis of images.

Five impairments were studied in the CableLabs test: (1) random noise, (2) phase noise, (3) chroma/luma delay, (4) microreflections, and (5) composite triple beat (CTB). These impairments are briefly described here (see Table 2.1). Random noise is the result of thermal motion of electrical charges and yields "snow" in the picture. Phase noise is the visual manifestation of instability of oscillators that are used in heterodyning devices to shift frequencies. Several of these systems are in the signal path from point of origination to the subscriber's screen. Phase noise usually appears as near-horizontal streaking of the picture. Chroma/luma delay results in the colored part of the image misregistering from the luminance part of the picture. The color is smeared and located in the wrong place, giving rise to a comparison with comic book drawings. Microreflections

Table 2.1 Suppression of Impairments for Various Subjective Reactions

Impairment	*Imperceptible*	*Perceptible, but not annoying*	*Slightly annoying*	*Annoying*
Random noise	53 dB	48–51 dB	45 dB	40 dB
Phase noise		90 dB	85 dB	
CTB		50 dB	40 dB	30 dB
Chroma/luma delay			300 ns	400 ns

cause delayed versions of the signal to add destructively or constructively. In many cases, the picture is blurred. In a few circumstances, the subjective sharpness of the picture may appear to improve even though the objective measurements may lead to a different conclusion. Composite triple beat (CTB) is the creation of new, undesirable frequency components due to the signal's passing through nonlinearities. The nonlinearities cause new frequencies to be created that interfere with the quality of the picture.

Table 2.1 summarizes the results. Quantitative data was not taken for microreflections and so is not included in the table. However, microreflections below 20 dB were not of consequence to most viewers. Below 10 dB, some pictures looked sharper, whereas others looked softer. The most important conclusion is that subscribers have become more critical and are likely to become even more so as they see more images delivered with digital television (DTV).

2.5.2 The W Curve and Direct Pickup (DPU)

In 1986, CBS Laboratories and the NAB tested subjective reaction of viewers to co-channel interference.[24] The desired signal was at broadcast channel 23 (visual carrier at 525.5 MHz) and consisted of three images: a gray flat field, a still image (of a boat), and moving video. The undesired signal was a series of sine waves at intervals of 500 kHz, starting at the frequency of the lower band edge of the channel. The ratio of the strength of the interfering sine wave to the strength of the visual carrier was adjusted until expert viewers agreed that the interference was "just perceptible." To be "just perceptible" at the lower band edge, the strength of the interfering sine wave needed to be attenuated so that it was only some 15 dB below the visual carrier strength. As the frequency of the interfering carrier was increased, more attenuation was required to bring it down to levels that were "just perceptible." Maximum attenuation, some 68 dB, below visual carrier strength was required at a frequency 250 kHz above the visual carrier. Less attenuation, about 55 dB, was required at the middle of the channel. Again more attenuation, in the low 60s, at the chroma carrier was required. And then, as the upper band edge was approached, less attenuation was required. These measurements yielded a W-shaped curve, as shown in Figure 2.23. The video content of the desired signal had some influence on the amount of attenuation required in the middle of the channel. Somewhat surprisingly, the flat field hid the interference better in the low to mid-frequencies, whereas the moving picture hid the interference better around the color subcarrier. Most surprising was the finding that the difference between the "just perceptible" level and unacceptable levels was just 2–3 dB.

Although co-channel interference is not generally a problem for cable television, direct pickup (DPU) interference is a major headache. In the case of a strong broadcast channel that is carried on the same (or close) frequency on cable, the DPU may result in a ghost or smearing of the image, since the

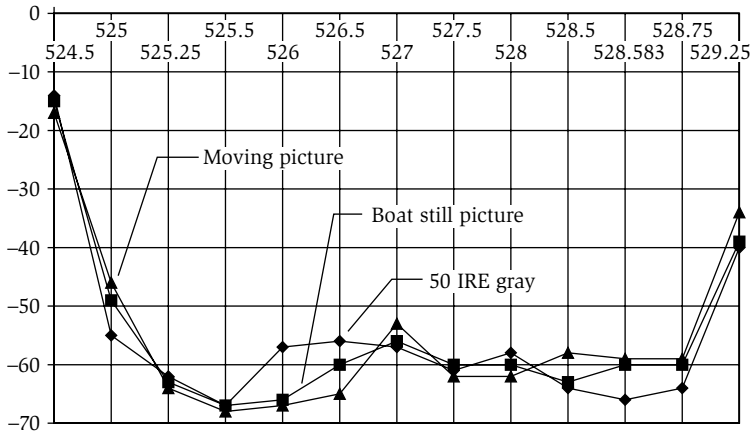


Figure 2.23 W curve: Level of "just perceptible" interfering carriers for three images.

propagation time through the cable is longer than through the air. In the case of other users of the spectrum, second-level-order distortion products, local UHF broadcasters, or cable systems using other than off-air frequencies, interfering patterns may appear on the screen. One manifestation of this is in the cable channel 19 area, where pager transmitters are authorized to operate. Direct pickup can occur in the cable plant or in equipment connected to the cable drop. Those DPU signals can propagate between products connected to the same drop and cause picture degradation. If the tuner of a TV or VCR is not adequately shielded, the DPU will cause ghosts and other interference. An occasional problem arises when subscribers connect FM radios to the cable. These products generally have almost no shielding on their tuners and may conduct DPU signals back into the cable system to interfere with reception by other devices. The W curve gives an indication of sensitivity to DPU at various frequencies.

In 1993, the 1986 CBS tests were repeated, with emphasis on their application to DPU. Since the TASO observations were found to no longer be applicable, it was reasonable to wonder whether consumers had become more sensitive to co-channel and DPU problems. The tests were done on three cable channels (6, 12, and 78), at four frequencies relative to visual carrier (+0.25, +0.75, +1.75, and +2.55 MHz), at two desired signal strengths (0 and +10 dBmV), and at two carrier-to-noise ratios (43 and 50 dB). Only the gray flat field was used for the desired signal. The 1993 tests generally confirmed the results of the 1986 CBS experiments. The W shape of the curve remained. Across the three channels, four frequencies, two desired signal levels, and two carrier-to-noise ratios, the average value of the sine wave strength attenuation relative to the carrier strength to achieve "just perceptible" impairment levels was 56 dB. Since these tests involve human subjects and a wide range of variables, they are difficult to repeat and obtain closely similar results.

2.6 Other Television Systems

The NTSC color system was historically the first to be adopted in the world. Because of the close economic relationships between the North American countries and Japan, all these countries adopted the same color television system, allowing products to be made and sold in these countries interchangeably, with only minor modifications due to differing channel frequencies.

Other countries adopted color television later.[25] As a result, they were able to take advantage of the experience of the NTSC countries and make modifications that better suited their priorities. In some cases, these priorities included the desire to protect domestic manufacturing by choosing standards that were different. Controlling the licensing of the intellectual property (patents) erected market-protecting barriers.

In the NTSC system, there are two distinct types of crosstalk between the luma and the chroma signals. Linear crosstalk is a consequence of luma signals (under some circumstances) having spectral components that lie in the chroma regions. The classic example of this is the spurious rainbow pattern that appears on a checkered or tweed jacket. The pattern produces frequencies that stimulate the chroma circuits. Conversely, the 3.58-MHz chroma signals are not completely eliminated from the luma circuits and appear as small moving dots, especially on sharp edges in highly saturated color areas. Slight nonlinearities cause the luminance and chrominance signals to modulate each other. In particular, the chroma phase and amplitude become a function of the signal's brightness. These phenomena, known as *differential phase* and *differential gain*, cause infidelities in color reproduction. As circuit technology has improved, these problems have become less severe. But in the first years, using vacuum tube circuits, great care and expense was required to keep these problems manageable.

The primary objective of European developments in color television was to minimize the impact of differential phase and differential gain and of other undisciplined phase shifts in the signal path. The principle behind the approach is to shift the phase on one of the chroma signals by 90° on adjacent lines. This approach gives rise to the technique's name: phase alternating at line (PAL) rate. In addition, the chroma reference burst also shifts 90° at the end of each line. The beneficial impact of this approach is that hue shifts on adjacent lines are in opposite directions. Two approaches can be used to take advantage of this. In the less expensive and simpler approach, it is left to the observer's eye to average the hues. This is quite effective for the small phase errors of modern equipment. However, larger phase errors give rise to horizontal lines sometimes called Hanover bars or Venetian blinds. The more expensive alternative is to use delay lines to electrically average the signals before display. While avoiding the artifacts mentioned, vertical chroma resolution is halved and additional expense results.

PAL also requires a relationship between the chroma subcarrier frequency and the line scan rate. However, because the phase of the chroma burst changes 90°

at the end of every line, the relationship must be that the subcarrier frequency is an odd multiple of the sum of one-quarter of the scan line frequency plus one-half the frame frequency. This is in contrast with NTSC's simpler relationship of an odd multiple of one-half of the horizontal scan frequency.

The PAL strategy of alternating phase on adjacent lines and the need to create a relationship with an odd multiple of one-quarter the horizontal scan frequency means that the chroma spectral lines are not centered in the luma spectral gaps. In the time domain, the averaging of brightnesses caused by the chroma signal is even less effective. The saving grace is that the chroma subcarrier frequency is almost a megahertz higher than in NTSC, and therefore the chroma dots are much smaller and less visible.

European television uses the UHF frequencies and thus has more bandwidth available for channelization. Whereas the North American and Japanese video baseband bandwidth is 4.2 MHz, the British have 5.5 MHz and most of Europe has 5.0 MHz. The French and some of the former Soviet bloc countries use 6.0 MHz. Additionally, most of the European countries utilize a 50-Hz power line frequency and thus have chosen a vertical monochrome frame rate of 50 frames per second. The consequence of the wider bandwidth and lower frame rate is higher vertical and horizontal resolution. These systems have 625 scan lines per frame, compared to the NTSC's 525 scan lines. The horizontal scan line rate is half of 50 times 625, or 15,625 Hz. However, the lower frame rate gives rise to flickering in bright areas of the image and to reduced temporal resolution. The PAL chroma subcarrier is at approximately 4.43 MHz (compared with NTSC's 3.58 MHz). A further benefit of the wider bandwidth is that both color signals in PAL have the same bandwidth, 1.3 MHz. The NTSC strategy of two different bandwidths for matrixed colors matching the human response to flesh tones is avoided. Simple color difference signals are used involving red minus luma and blue minus luma.

The French SECAM system is more complex and uses frequency modulation for the two color difference signals. These two signals are alternated on adjacent lines, yielding half vertical color resolution.

2.7 Summary

The fundamentals of analog television are covered in a manner intended to build an understanding around which to better appreciate cable television technology. We presented the basic concepts behind the structure of the signal and its modulation. This allows the requirements on the signal processing implemented in cable to be developed in later chapters. Bandwidth compression is important in analog television. The interlaced scanning system and vestigial sideband modulation are both important bandwidth compression techniques. When color is added to the signal, the careful matching of the signal to the human color visual response saves bandwidth.

We surveyed the consequences of the technical compromises forced by the need to have affordable receivers in the early days of television. We explored the important concept of compatibility, demonstrating that complete compatibility is not necessary. Some modest degree of degradation of the performance of earlier products is acceptable in order to advance the performance of new products. A balance is struck.

The story of NTSC is replete with technical compromise. The goals have always been to match image and sound quality to the human who is expected to pay for the hardware and the services. A good compromise and a good bargain have been attained.

Endnotes

1. Donald G. Fink, ed., *Television Standards and Practice: Selected Papers from the Proceedings of the National Television System Committee and Its Panels*. New York: McGraw-Hill, 1943.
2. David E. Fisher and Marshall Jon Fisher, *Tube, The Invention of Television*. The Sloan Technology Series. Washington, DC.: Counterpoint Publishers, 1996.
3. Ibid.
4. Bernard Grob, *Basic Television*, 3rd ed. New York: McGraw-Hill, 1964.
5. K. Blair Benson and Jerry Whitaker, *Television Engineering Handbook*, rev. ed. New York: McGraw-Hill, 1992.
6. Bernard Grob, op. cit.
7. Mischa Schwartz, *Information Transmission, Modulation, and Noise*, 2nd ed. New York: McGraw-Hill, 1970.
8. K. Blair Benson and Jerry Whitaker, op. cit.
9. B. P. Lathi, *Communication Systems*. New York: Wiley, 1968.
10. Mischa Schwartz, op. cit.
11. Donald G. Fink, *Principles of Television Engineering*. New York: McGraw-Hill, 1940.
12. *Proceedings of the IRE, Color Television Issue*, Vol. 39, No. 10, October 1951. The Institute of Radio Engineers, New York.
13. *Proceedings of the IRE, Second Color Television Issue*, Vol. 42, No. 1, January 1954. The Institute of Radio Engineers, New York.
14. *Color TV Training Manual*, 3rd ed. Indianapolis, IN: Howard W. Sams, 1970.
15. A. M. Morrell, H. B. Law, E. G. Ramberg, and E. W. Herold, *Color Television Picture Tubes*. New York: Academic Press, 1974.
16. Keith Jack, *Video Demystified*. San Diego, CA: HighText Publications, 1996.

17. Theodore S. Rzeszewski, *Television Technology Today*. Piscataway, NJ: IEEE Press, 1985.
18. Special Issue on Consumer Text Display Systems, *IEEE Transactions on Consumer Electronics*, Vol. CE-25, No. 3, July 1979.
19. C. E. Dean, Measurements of the Effects of Interference in Television Reception. *Proceedings of the IRE*, Institute of Radio Engineers, New York, June 1960.
20. G. Chouinard, W. A. Whyte Jr., A. A. Goldberg, and B. L. Jones, Experimental Results Supporting the Determination of Service Quality Objectives for DBS Systems. *IEEE Journal on Selected Areas in Communications*, Vol. SAC-3, January 1985.
21. B. L. Jones and J. A. Turner, Subjective Assessment of Cable Impairments on Television Picture Quality, *1991 NCTA Technical Papers*, p. 92. Washington, DC: NCTA.
22. Method for the Subjective Assessment of the Quality of Television Pictures, Recommendation 500-4, *Recommendations and Reports of the CCIR*, Vol. XI, Part 1. Dubrovnik, Croatia: CCIR, 1986.
23. B. L. Jones and L. E. Marks, Picture Quality Assessment: A Comparison of Ratio and Ordinal Scales. *SMPTE J*, Vol. 94, No. 12, December 1985.
24. G. Chouinard et al., op. cit.
25. Geoffrey H. Hutson, *Colour Television Theory, PAL-System Principles and Receiver Circuitry*. London: McGraw-Hill, 1971.

Other References of Interest

Donald Christiansen, *Electronics Engineers' Handbook*, 4th ed. New York: McGraw-Hill, 1996.

Margaret Craig, *Television Measurements, NTSC Systems*. Beaverton, OR: Textronix, 1990.

Jeffrey L. Thomas, *Cable Television Proof-of-Performance*. Englewood Cliffs, NJ: Prentice Hall, 1995.

Ted Rzeszewski, *Color Television*. Piscataway, NJ: IEEE Press, 1983.

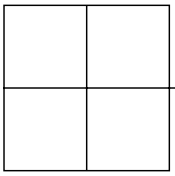
Joel Brinkley, *Defining Vision, The Battle for the Future of Television*. New York: Harcourt Brace, 1997.

J. Allnatt, *Transmitted Picture Assessment*. New York: Wiley, 1983.

B. L. Jones, Subjective Assessment of Cable Impairments on Television Picture Quality, *1992 NCTA Technical Papers*, pp. 1–25. Washington, DC: NCTA.

NCTA Technical Papers, yearly, NCTA, Science and Technology Department, Washington, DC.

IEEE Transactions on Consumer Electronics, quarterly, Institute of Electrical and Electronic Engineers, Piscataway, NJ.



Chapter 3
Digitally Compressed Television

3.1 Introduction

This chapter introduces digital compression, transport, and transmission for television signals, particularly as they relate to cable systems. However, it is worth taking a moment to consider why the cable industry has transformed itself to embrace digital technology. After all, in the mid-1990s the cable industry was starting a hybrid fiber-coax (HFC) upgrade program to expand channel capacity to greater than 100 analog channels. Why then was digital technology adopted so rapidly in the late 1990s after many HFC upgrades had already been completed?

Digital compression was first applied to moving images in video conferencing applications.[1] From there, interest in its application to television signals started with early satellite systems (such as SkyPix and PrimeStar). However, the essential push may have been provided by high-definition television (HDTV).[2] An HDTV signal contains about six times as much information as a standard-definition television (SDTV) signal and consumes at least 12 MHz of spectrum in analog format. This fact makes digital compression a necessity for HDTV. Once the investment had been made to digitally compress an HDTV signal, the same technology was naturally applied to SDTV. The first application was in direct broadcast satellite (DBS) systems — digital compression was a perfect fit for a new system that was constrained by expensive, and limited, satellite transponder capacity. Digital transmission also reduced the required signal-to-noise ratio, allowing for a smaller satellite dish (on the side of the home).

With digital compression, DBS systems could offer enough channels to pose a competitive threat to cable systems, especially those that had not done an HFC upgrade. The range of programming continued to expand, and suddenly even 100 analog channels wasn't enough to carry them all. Moreover, the DBS industry began to actively market its "digital picture and sound quality." As a result, the cable industry turned its attention to digital. The new digital tier allowed cable to lift its per-subscriber revenue by adding new premium subscription content, and the pay-multiplex was created by HBO. In addition, digital pay-per-view greatly expanded choice for the customer. Finally, digital compression allowed an upgraded system to offer video-on-demand, something DBS could not offer.

Traps and analog scrambling are relatively easy to circumvent, and signal theft has become a significant issue in analog systems. Digital technology enables a more secure conditional access mechanism based on *cryptography* (see Chapter 21).

This chapter will focus on the methods employed by the satellite, cable, and broadcast systems, all of which use MPEG-2 video compression. In addition, they are all based on a one-way, time division multiplexed approach known as *MPEG-2 transport*, although there is growing interest in alternative methods that are more suitable to Internet Protocol (IP) networks under the umbrella term *streaming media*.

We will consider video compression and then audio compression before delving into multiplexing and data transport of compressed digital content over existing satellite, cable, and broadcast systems. First, what is the history of broadcast digital television?

3.2 Broadcast Digital Television

The NTSC television system was created at a time when electronics were expensive and signal processing difficult and limited. The programs were few, and they were relatively expensive to produce. An audience was lacking. A great many compromises were made to allow cost-effective receivers. Without affordable receivers, there would be no point in having efficient use of the spectrum. The priority was rapid penetration of the marketplace and the generation of demand for programming and receivers.

The situation is now quite the opposite. The scarce resource is spectrum. Electronics and electronic processing are affordable and on a predictable path of further cost reduction (according to Moore's law; see Section 3.2.2). The goal now is to create systems that efficiently utilize the spectrum and that will last for several decades without becoming painfully obsolete. The first step is to study the redundancies in television images that can be eliminated to lighten the spectrum burden.

Interest in digital television goes back several decades. Long before digital television was practical for consumer use, it held interest for military and professional applications. Digital signals have some important advantages:

- The ability to transmit signals over arbitrarily long distances (using repeaters if necessary; see section 3.2.2)
- The ability to store the signals without degradation
- The ability to process signals for security and other purposes
- The ability to remove unnecessary redundancy to increase transmission and storage efficiency
- The ability to add appropriate coding and redundancy so that signals can survive harsh transmission environments

3.2.1 High-Definition Television (HDTV)

The big push for consumer digital television came as a consequence of the search for a practical high-definition television (HDTV) system. HDTV was defined as having twice the horizontal resolution, twice the vertical resolution, a wide picture with a ratio of 16 units of width for every 9 units of height, no visible artifacts at reasonable viewing distances, and compact disc–quality sound. Before any processing, the analog signal coming from an HDTV camera could consist of 30 MHz of red information, 30 MHz of green information, and 30 MHz of blue information. Almost 100 MHz of analog information is involved. To convert it to digital form, it is first sampled at twice the highest frequency (according to the Nyquist theorem; see Section 3.3.2). Then each sample is represented by a byte of data. More than a gigabit per second of data transmission is required. It can be appreciated why all the early HDTV proposals were analog! Methods of reducing this bandwidth appetite had to be found. The work on HDTV began in Japan in 1970 using analog transmission technology coupled with significant digital processing at both the point of origination and the receiver. The Japanese proposal was called multiple sub-Nyquist sampling encoding (MUSE). It applied many of the techniques used to minimize NTSC bandwidth. The goal was to match the HDTV system to the human visual response, since there is no need to transmit what the eye does not see.

When the FCC decided to pursue HDTV in the United States, it laid down a number of challenges.[3] The first challenge was for a "compatible" HDTV system that would not make current NTSC receivers obsolete. The second major challenge was to limit the original NTSC and the new HDTV signals to no more than two 6-MHz slots in the television spectrum. A further restriction required the new signals to avoid undue harm to existing NTSC transmissions. After an extended search for a compatible method of creating HDTV, it became clear that all methods proposed used the original NTSC signal plus in-band and out-of-band "helper signals." All these resources were required to create the compatible signal, and two 6-MHz bands were consumed for each HDTV signal. This approach meant that NTSC would always have to be supported. It also meant that 12 MHz was committed for each HDTV signal. If there were ever to be a transition away from broadcast NTSC, this approach would have to be abandoned.

Zenith Electronics Corporation first broke ranks with the analog proponents by proposing a hybrid system that continued the transmission of the high frequencies of the image in analog form but converted the lower frequencies to a digitized form. This hybrid approach seemed to use the best of both worlds. It recognized that most of the energy in an NTSC signal is in its low frequencies, which include the synchronization pulses. In NTSC, the sync pulses have the highest energy because they are separated from the video signal by amplitude discrimination. By digitizing the low frequencies, the majority of their power consumption was eliminated, yet the burden on the digital circuits was relaxed because only relatively low frequencies were processed. The high frequencies

remained analog and contributed little to the power requirements. The lower-data-rate digital signals would also be less susceptible to multipath transmission effects. The remaining problem was that this approach was no longer compatible with existing NTSC receivers. This problem was solved by allowing compatibility to include simulcasting. That is, both the hybrid signal and the NTSC signal would carry the same programming, but at two different resolutions and at two different frequencies. This would preserve the utility of older receivers. Moreover, since no successful system that put both NTSC and HDTV into the same 6 MHz had been proposed, two 6-MHz channels would still be required, as in all the other proposed systems.

This approach had a number of major advantages. The low-power HDTV signal was tailored to cause minimum co-channel interference in adjacent locations using the same frequencies. Also, since the NTSC signal was not needed as a component of the HDTV signal, it could eventually be abandoned. The NTSC channel could then be reallocated to other purposes. Even before that happened, the requirement for simulcasting could be relaxed based on policy rather than technological constraints. Via this step-by-step process, compatibility was abandoned for the first time in broadcast television. (The noncompatible CBS color system, though temporarily the official system, did not achieve commercial success before it was replaced with the RCA/NTSC compatible color system.)

However, cable and satellite didn't have the same requirement for terrestrial compatibility. In fact, the lack of any viable analog HDTV solution for satellite led General Instrument's DigiCipher division to develop their purely digital approach. (At the same time the "clean sheet of paper" approach was also being advocated in some cable circles.) Shortly thereafter, General Instrument proposed an all-digital solution for terrestrial HDTV broadcast. Quickly, all serious proponents, with the exception of the Japanese MUSE system, converted to all-digital designs. The committee, charged with selecting a winner, found that it could not come to a conclusion. The technical issues were too complex, and the political issues were overwhelming. At the time a decision was to be made, all the proposed systems made unacceptable pictures. The result was an elaborate ruse to score all the systems as acceptable under the condition that a "grand alliance" be formed that allowed the proponents themselves to hammer out a single system. This allowed the political battles to go on behind closed doors under the guise of selecting the "best of the best" for a single proposal to the FCC.

Since that time, the television and computer industries have jointly created standardized toolkits for audio and video compression. The MPEG-2 video compression standard[4] provides a "toolbox" of techniques that can be selected according to the nature of an application. The MPEG-2 systems standard defines how multiple audio and video channels can be multiplexed together into a single digital bitstream. Finally, the broadcast and cable industries have developed modulation and transmission standards to reliably deliver the digital bitstream over an analog channel.

Using these standards, a single HDTV program can now be transmitted within the analog broadcast TV channel assignment of 6 MHz rather than the tens of megahertz once thought necessary. In the case of cable's well-behaved spectrum, double the data transmission rate is possible; two to three HDTV signals can be carried in 6 MHz.

The same technology that makes HDTV possible in 12–18 Mbps also enables very efficient compression of standard-definition television. Excellent results are now being obtained at data rates of 2–3 Mb/s using MPEG-2 video compression. Since the cable transmission rate is approximately 38 Mb/s (in 6 MHz), 12–18 programs can be carried in the same spectrum that previously carried one.

It is expected that the transmission of standard analog NTSC will occur for many years before a transition to digital television is completed. The British experience with vertically polarized, 405-line monochrome television is instructive. It took some 50 years to abandon that system, even with a very small number of receivers. A minority of opinion holds that the U.S. transition may never be completed because of the vast installed base of analog receivers (currently more than 250 million color receivers and more than 150 million VCRs) and the political problems associated with making consumers' "investment" in products obsolete. Low-income viewers and the economically challenged would be especially hard hit. It is difficult simultaneously to be a supporter of "free television" and to require the poor to purchase new, expensive television sets (not to mention that advertiser-supported television requires a large number of subscribers to sustain itself). Of course, the new digital receivers will rapidly come down in price, but the analog receivers will also continue to decline in price. The shutdown of analog television will not be easy.

3.2.2 Digital Signal Processing

The human sensory systems for images and sounds are analog. Images and sounds start as analog phenomena. To be enjoyed by humans, the phenomena must be converted to signals, and the signals must eventually be displayed as analog stimuli for eyes and ears to enjoy. Unfortunately, as analog signals are transmitted over long distances, they encounter noise, distortion, and interfering signals that degrade the quality of the images and sounds, so they become unpleasant and, eventually, unusable. If the analog signals are converted to digital signals, an arbitrarily small degree of noise is introduced in the conversion process, but all subsequent degradation of the signal can be avoided using practical and well-understood techniques.

The advantages of baseband (unmodulated) digital signals include the ability to completely regenerate the signal, thereby preventing the accumulation of noise and distortion, and the ability to apply computational techniques for the purposes of error detection and correction and redundancy reduction. Redundancy reduction is important because it saves data storage space and transmission time and bandwidth.

Advantages of binary numbers include that they have two states, which can be represented by simple and inexpensive circuits, and that the impact of electrical noise and distortion can be minimized or even eliminated by proper design. A circuit element, such as a transistor, that processes an analog signal must faithfully reproduce all values of the signal while adding a minimum of distortion and noise. If many such circuit elements process an analog signal, their individual contributions of noise and distortion accumulate, causing substantial signal degradation. On the other hand, a binary circuit can have two well-defined states, on and off, which are easily distinguishable. The on state may represent the binary number 1, whereas the off state can represent the binary number 0. (The opposite choice is equally valid.) The important point is that if the circuit element is *mostly off* but not completely off, it will not be confused with the on state. Similarly, if the circuit element is *mostly on* but not completely on, it will not be confused with the off state. Thus imperfect performance of the circuit can still faithfully represent the binary values. Only when the on state approaches half of the assigned value or when the off state is almost halfway to the on condition can confusion result. If this degree of deficient performance is avoided, the two states can be discriminated and the signal perfectly resolved (see Figure 3.1). If, as the signal is transmitted, it suffers some noise and distortion degradation, it can still be perfectly recovered as long as the two states — the on state representing a binary 1 and the off state representing a binary 0 — can be reliably discriminated. Eventually, sufficient noise and distortion will accumulate so that the two states become confused. If the system is designed so that the signal is regenerated (and any errors are corrected) prior to this destructive level of degradation, a fresh binary signal can be substituted for the degraded signal, and all the damage caused by noise and distortion can be completely undone. *Repeaters* can perform this operation an arbitrary number of times, allowing error-free communications over arbitrarily long distances. This is something that cannot be accomplished with analog signals.

A further advantage of digital signals is that the art of circuit design has progressed to where the transistors that process digital signals are very small and very inexpensive. Gordon Moore, one of the founders of Intel Corporation, observed that approximately every 12–18 months, the number of digital transistors that can be afforded for a given price doubles. Alternatively, the cost of a given number of digital transistors approximately halves in that period. This process has been continuing for decades and appears likely to continue for some time to come. As an example of this phenomenon, the first personal computers introduced in the early 1980s used an Intel-brand integrated circuit (IC) that had 30,000 digital transistors. The Pentium computer ICs of the mid-1990s have more than 5 million digital transistors. Tens of millions of digital transistors are now being placed in consumer products at affordable prices. The same experience has not been enjoyed by analog circuits, because they must faithfully process the infinite range of values of analog signals. That severe constraint has prevented analog circuits from progressing as fast or as far in complexity and cost reduction.

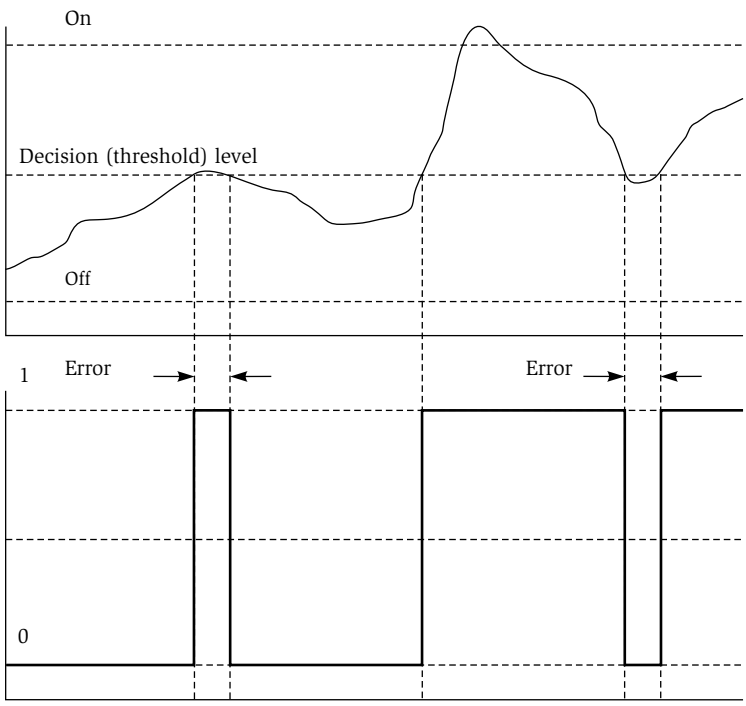


Figure 3.1 Impact of excessive noise.

An additional advantage of digital signals is that they can be mathematically manipulated in very complex ways that lead to methods of determining whether transmission errors have occurred and even of correcting some of those errors. Note that there are only two possible types of errors: a binary 1 symbol may be damaged and converted into a binary 0 symbol, and a binary 0 symbol may be damaged and converted into a binary 1 symbol. There are no other alternatives in a binary system. As an example, one common method of identifying an error condition is called *parity detection*. Binary symbols are grouped into clusters of seven, and an eighth symbol is appended depending on whether the previous seven have an even or an odd number of 1 symbols. If the appended symbol is such that an even number of 1 symbols is always present in each valid group of eight symbols, the method is called *even parity*. A single transmission error will convert a 1 to a 0 or a 0 to a 1, and the result will be an odd number of 1 symbols. The system can detect this error. If two errors occur, a much less probable event, the system will be fooled and think that no error has occurred, because there will still be an even number of 1s. However, if three errors occur, the damage will again be detected. The ability to detect certain error conditions has come at the

price of an appended symbol that takes up transmission time and that requires additional circuits to process the signal at both the transmission and receiving ends. Forward error correction (FEC) algorithms can detect multiple errors and even compute the original correct signal. These more complex methods increase the amount of additional, nondata symbols and are said to have increased *overhead*. In addition, additional processing complexity is required at both the sending and receiving ends of the transmission path.

3.3 Digital Video Compression

As we have seen in Section 3.2, uncompressed digital video requires a very high data rate; in fact, these rates are so high that it would be uneconomical to transmit or store uncompressed digital video. Therefore data compression is required so that the video is still of acceptable quality while being described by many fewer bits. This type of compression is known as *lossy* compression, because some information is discarded during the encoding process. "Acceptable quality" is defined by customer expectations and business conditions and is, by nature, always something of a compromise. For both operators and programmers, more is better; the more aggressive the compression, the more channels for a given bandwidth. Many television viewers seem to be very forgiving of a softer picture (that is, a loss of resolution) as long as the picture remains noise free. Nevertheless, more sophisticated viewers have certainly shown their preference for better resolution, as is demonstrated by the phenomenal success of DVD. This trend may continue as large-screen, high-resolution displays become more affordable, and it bodes well for HDTV adoption.

The first large-scale application of digital video compression was, and still is, broadcast television services (in the order of satellite, cable, and terrestrial broadcast). These applications all require a standard coding mechanism that is understood by all decoders in the marketplace. The familiar MPEG standard is the best example of this. It also allowed for volume production of MPEG-2 decoder chips so that they became inexpensive enough to put in the set-top terminal (STT) and, eventually, the television set.

A second application is the digital versatile (originally video) disk (DVD). Just as in broadcast television services, the decoder in every DVD player had to perform identically; once again, the MPEG-2 video format was adopted.

Finally, video on demand (VOD), which is now being aggressively deployed by many cable operators, also uses MPEG-2 standards. In this case, MPEG-2 is a natural choice, because VOD leverages the existing investment in the STT.

A very different application, called *streaming media*, has emerged with the Internet. In streaming media, the transfer of video is typically a 1-to-1 transaction, making it analogous to VOD. However, the Internet environment allows the coding format to be specified for each transaction, even allowing a software decoder to be downloaded just before it is needed. Moreover, the

ubiquitous personal computer now provides enough raw central processing unit (CPU) performance and memory to allow the decoding to be done without special-purpose silicon. These properties make the Internet a perfect incubator for new compression algorithms.

3.3.1 Principles of Video Compression

There are many excellent texts that describe the principles of digital video compression.[5–7] The reader is referred to these for a more detailed treatment.

Digital video compression can achieve a dramatic reduction in bit rate; a factor of 50:1 is quite achievable for entertainment-quality video. There are two main reasons for this.

1. A typical video stream contains a tremendous amount of spatial and temporal redundancy.
2. The human eye–brain system is relatively insensitive to certain impairments in the displayed result.

In the first case, recall that NTSC video is refreshed 60 times per second (and PAL video 50 times per second) to avoid the perception of flicker. There are many segments when very little motion has to be conveyed, and one field is much like the one before it (temporal redundancy). Fundamentally, if the changes could be sent, rather than completely refreshing the field, then a lot of data transmission could be saved.

Second, the human eye–brain system is quite forgiving when it comes to the fidelity of video rendition. For example, the eye's rod cells are very sensitive to the total intensity of light (luminance), whereas the eye's cone cells are less sensitive to color differences, particularly for certain shades. The eye is also less sensitive to color (or chrominance) detail. NTSC takes advantage of this to dramatically limit the bandwidth of the chrominance (I and Q) signals, as explained in Chapter 2. Finally, the eye–brain system is even less sensitive to detail if there is significant motion in that part of the scene.

These discoveries about the human eye–brain system have been researched thoroughly and codified into a perceptual model that can be used to reduce fidelity in those areas where it is least noticeable while maintaining fidelity where the eye is most discriminating. Nevertheless, much advanced compression work transcends perceptual modeling, becoming quite empirical. This explains the lack of good, objective video-quality measurement techniques based on perceptual modeling.

3.3.2 Coding of Still Images

We will start by discussing the compression of a still image and then extend the model to deal with moving images. Video can be represented as a series of still

images. In fact, we must periodically code an entire frame as a still image (an intracoded frame, or I frame) because this provides the property of random access. An I frame contains all of the information needed to decode the frame within the frame itself; it is not dependent on adjacent frames, and it serves as a reference for these other frame types. Random access is very important for broadcast applications; when the viewer is switching from one channel to the next, it must be possible to reinitialize the decoding process within a short period of time (less than a second).

In both still- and moving-image compression, a number of key steps are commonly applied:

1. Sampling and preprocessing
2. Block formation
3. Transform coding
4. Quantization
5. Run-length encoding
6. Entropy coding

(This section describes how the fundamental principles of compression are implemented by MPEG-2 video compression of a standard-definition TV picture, because it is so widely used. Subsequent sections will expand the scope to include discussion of HDTV, MPEG-4 video, and alternative approaches to video encoding.)

Sampling and Preprocessing

To convert an analog signal to a digital signal, it is first sampled in time. This can be accomplished by multiplying the analog waveform by a train of evenly spaced narrow pulses that alternate between zero volts and a constant, convenient voltage (for example, 1 volt) (see Figure 3.2). The information science theorist Harry Nyquist of Bell Telephone Laboratories proved that if a signal is sampled at least twice as frequently as the maximum signal frequency it contains, it can be perfectly recovered with no loss of information.[8,9] This principle is called the Nyquist theorem. Sampled signals are still analog signals, because they can take on any value; they are just time quantized.

Nyquist proved that if the sampled signal is low-pass filtered, the original signal is recovered, with just a delay caused by the propagation time through the circuits (see Figure 3.3). If each time sample's strength is then measured (for example, with a digital voltmeter) and the resulting measurement represented by a number of limited precision, the sampled analog signal will be converted into a sequence of data (see Figure 3.4). Limited-precision numbers have a fixed number of decimal places. The uncertainty in precision of the number is determined by the value of its last decimal place. The information to be transmitted is

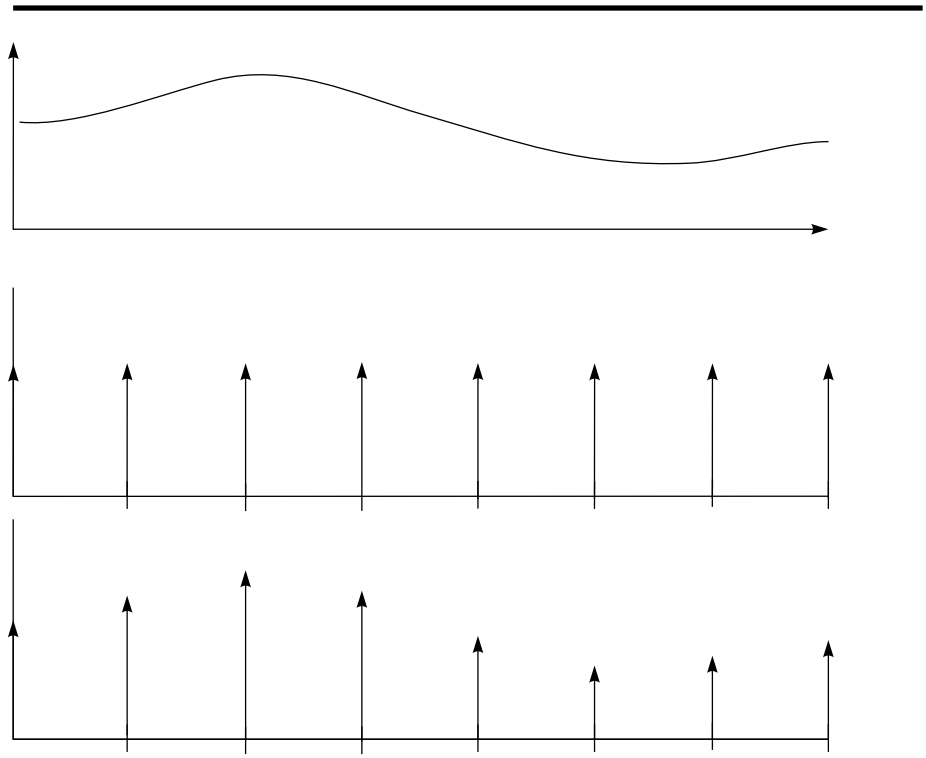


Figure 3.2 Time sampling an analog waveform.

no longer the original analog signal or its time-sampled version (which can take on any value), but rather another signal that conveys the limited-precision numbers describing the strength of the original signal samples. The representation of the signal by a limited-precision number introduces an error that can be considered as a type of noise called *quantization noise*. The amount of quantization noise can be made arbitrarily small by using arbitrarily higher-precision numbers, but it can never be reduced to zero.

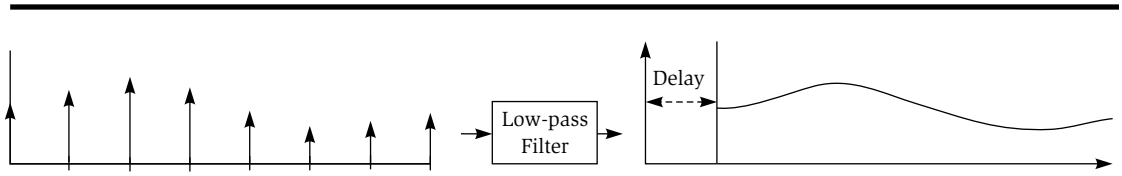


Figure 3.3 Recovery of a time-sampled waveform.

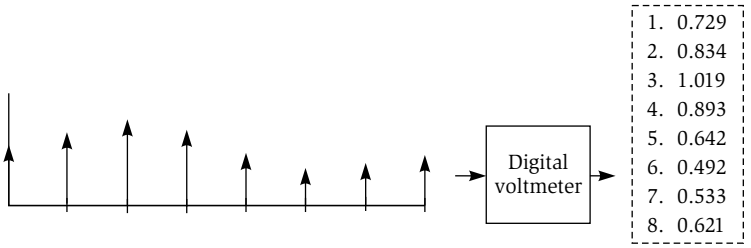


Figure 3.4 Quantization of a time-sampled waveform.

Video can be sampled in many different ways — ironically, sampling to achieve studio-quality transfers also tends to generate the highest data rate and is therefore not the best approach for low-bit-rate encoding. However, the "cleaner" the video, the better, because any noise in the original signal is random, additional information that is not easy to compress. For this reason, it is common practice to low-pass filter the video signals before sampling, in an attempt to reduce any high-frequency noise components. Achieving good noise reduction without unduly reducing resolution has been a subject of considerable research and development among encoder manufacturers.

As described in the previous chapter, color television is based on the discovery that the human eye responds to three colors and mixes them to produce essentially all other colors. Color television requires three signals. The three signals start at the camera and end at the picture tube as red, green, and blue. In between, they can be transformed into other formats that are more convenient for processing. In digital television, the color information is sent as two color difference signals, or *components*. The first component, C_b, is the difference between the luminance value and blue in that part of the picture. The second component, C_r, is the difference between the luminance value and red in that part of the picture. Since the human eye is less sensitive to detail in color information than in luminance (brightness), the color information can be sent with less detail.

As mentioned already, to be properly reproduced, an analog signal must be sampled at a rate at least twice that of its highest frequency. That theoretically minimum rate requires perfect (nonrealizable) filters at the receiving end. To use practical filters, an even higher sampling rate is required. The most common sampling rates chosen for use in digital television with resolution comparable to NTSC are 13.5 MHz for the luminance signal and half that rate (6.75 MHz) for each of the two components, for a total of 27 megasamples per second.

Interestingly, the use of compression allows the program bandwidth and the transmitted bandwidth to be very different. A modern television camera for NTSC actually produces video frequencies beyond what can be transmitted in the 6-MHz broadcast channel. Those frequencies are just thrown away in analog

practice. In compressed television, some of them may be used. The 13.5-MHz sampling rate supports a luminance bandwidth of 5.75 MHz and supports 2.75 MHz in each of the color difference channels. The NTSC technique of choosing color axes that favor the human visual system and limiting one to 0.5 MHz while allowing the other to go to 1.5 MHz is not necessary. Compression allows both color difference signals to have the same 2.75-MHz bandwidth.

It was determined by experiment that a minimum of eight bits per sample are required to minimize visible artifacts due to the quantization of the data. In some applications, 10 bits per sample are used. In digital television systems using 27 megasamples per second and eight bits per sample, uncompressed transmission requires 215 Mb/s for the video data. Of course, audio must be digitized as well. Using our bandwidth rule of thumb, something over 100 MHz of bandwidth would be required. This is not a problem internal to systems that process a single channel or for transmission of a few channels over fiber-optic links. However, this kind of bandwidth consumption is not feasible for broadcast or for multichannel environments, where the number of channels is important to subscriber appeal. There is a clear need for methods of reducing the bit rate to acceptable levels.

Because NTSC has fewer lines (525) and a higher frame rate (29.97), while PAL has more lines (625) and a lower frame rate (25), both forms of analog video can be conveniently sampled at the same rate to create the basis for a worldwide digital video rate hierarchy.[10] Sampling of an analog video source according to ITU-T Recommendation BT.601-5[11] reduces the chrominance data rate by sampling at 4:2:2.* The meaning of these numbers is:

4. The luminance component (Y) is sampled at 13.5 MHz.
2. The first chrominance component (C_r) is sampled at 6.75 MHz.
2. The second chrominance component (C_b) is sampled at 6.75 MHz.

Here, Y, C_b,C_r are the digital equivalents of the luminance, I, and Q components of an analog picture. Just as in the analog case, the color difference signals are manipulated to provide the best fidelity where it is most critical for the human eye.

In both NTSC and PAL, the active picture region is 720 pixels wide, but the first eight and the last eight are usually ignored, so only the middle 704 pixels are encoded. In the vertical direction, after VBI lines are discarded, there are 483 active lines in an NTSC picture, though only 480 lines are encoded. For PAL, depending on the version, there are between 575 and 587 active lines; only 576 are encoded.

---

* Originally the number 4 originated as being 4 times the color subcarrier frequency — for NTSC this is 4×3.58 MHz (or 14.3 MHz); however, a slightly lower rate of 13.5 MHz was chosen by ITU-R BT.601-5 for compatibility with PAL.

Block Formation

The next step in the process is to represent the video as a two-dimensional array of pixels so that it can be processed in the spatial domain. To do this, the digital video stream is deinterlaced and clocked into a frame buffer. (For images with high-motion content, this can cause a double image because there is movement between the two frames. We'll discuss ways to deal with this problem later; see Section 3.3.3, Coding of Moving Images.)

The frame buffer is divided into blocks typically 8×8 pixels because this is a convenient size for transform coding (see next subsection). Before this is done, however, the chrominance component is subsampled again, this time in the vertical direction, by averaging the samples on adjacent rows in the frame buffer. Sampling at 6.75 MHz and then averaging between rows reduces the chrominance resolution by a factor of 4 (compared to luminance). This resolution is termed 4:2:0. Therefore, for every four Y (luminance) blocks, there is one C_b block and one C_r block, as shown in Figure 3.5. (This grouping of blocks is called a macroblock in MPEG parlance.)

The vital statistics for standard-definition digital video are summarized in Table 3.1.

Transform Coding

As discussed in the previous chapter, waveforms that are cyclically repetitive can be decomposed into a series of sine and cosine waves of varying amplitudes. Equations (3.1), (3.2), and (3.3) display the familiar Fourier series equations discussed in the last chapter.[12]

$$f(t) = a_0 + \sum_{n=1}^{\infty} [a_n \cos(n\omega_0 t) + b_n \sin(n\omega_0 t)] \qquad (t_0 < t < t_0 + T) \qquad (3.1)$$

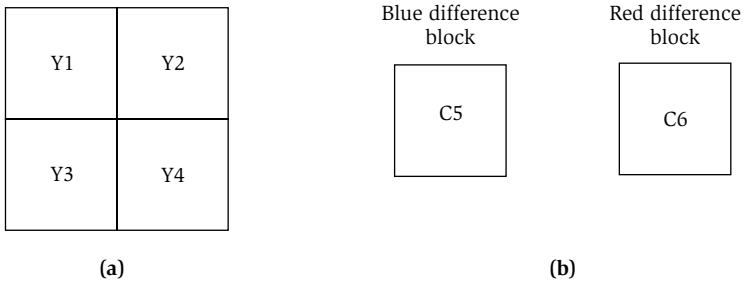


Figure 3.5 Order of transmission. (a) Luminance blocks. (b) Color difference blocks.

Table 3.1 Vital Statistics for Standard-Definition Digital Video

ITU-T BT.601-5/MPEG-2 main level	*NTSC*	*PAL*
Frame rate	29.97 Hz	25 Hz
Sampling rate:		
Luminance	13.5 MHz	13.5 MHz
Chrominance	6.75 MHz	6.75 MHz
Samples per line	858	864
Active picture samples	720	720
Encoded pixels per line	704	720
Vertical lines	525	625
Active lines	483	575–587
Encoded lines	480	576
Frame resolution (pixels)	704×480	704×576
Total macroblocks per frame	1,320	1,584
Total blocks per frame	7,920	9,504

Equation (3.1) is the trigonometric Fourier series representation of time waveform $f(t)$ over time interval $(t_0, t_0 + T)$. The various constants a_n and b_n are given by

$$a_n = \frac{\int_{t_0}^{(t_0+T)} f(t)\cos(n\omega_0 t)dt}{\int_{t_0}^{(t_0+T)} \cos^2(n\omega_0 t)dt} \tag{3.2}$$

and

$$b_n = \frac{\int_{t_0}^{(t_0+T)} f(t)\sin(n\omega_0 t)dt}{\int_{t_0}^{(t_0+T)} \sin^2(n\omega_0 t)dt} \tag{3.3}$$

If a waveform is not periodic, such as shown in Figure 3.6(a), it is possible to define a periodic version, as shown in Figure 3.6(b).

The Fourier series in Equations (3.1), (3.2), and (3.3) will be valid over the interval from time t_0 to time $t_0 + T$ but incorrect elsewhere. This is not a problem if we simply reapply the technique of forming periodic extensions for each piece of the waveform. We can then process the signal one piece at a time and load the results in memory.

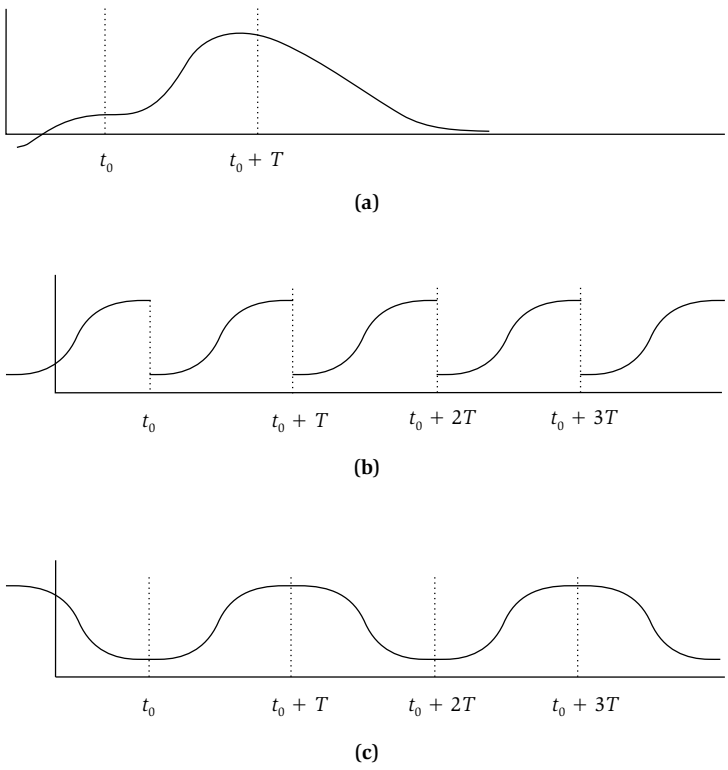


Figure 3.6 Periodic construction. (a) Nonperiodic waveform. (b) Periodic extension. (c) Symmetric periodic extension.

If a waveform has symmetry about the zero time point, then only cosines are needed to represent it, and all the b_n terms become zero. These cosines occur at multiples — or harmonics — of the fundamental repeat frequency of the waveform.

Figure 3.6(c) shows that the periodic extension technique can be applied in a manner that makes the signal symmetric so that its b_n terms will become zero. Again, the Fourier series is valid over only the interval from time t_0 to time $t_0 + T$ but incorrect elsewhere.

The amplitude values of these cosine waves constitute the cosine Fourier series representation of the waveform. The Fourier representation simplifies to just Equations (3.4) and (3.5).

$$f(t) = a_0 + \sum_{n=1}^{\infty}(a_n \cos n\omega_0 t) \qquad (t_0 < t < t_0 + T) \tag{3.4}$$

$$a_n = \frac{\int_{t_0}^{(t_0+T)} f(t)\cos(n\omega_0 t)dt}{\int_{t_0}^{(t_0+T)} \cos^2(n\omega_0 t)dt} \tag{3.5}$$

We are going to sample the video waveform according to the Nyquist theorem (as discussed in Section 3.3.2). Samples will be taken at more than twice the video waveform's highest frequency. Consequently, the video waveform will no longer be continuous. It will exist only at a finite number of discrete points in time. Therefore the integrals of Equation (3.5) become simple summations because there is only a finite number of sample points to add up rather than a continuous set of values to integrate. Since we are dealing with a bandwidth-limited signal, the summation of Equation (3.4) does not go to infinity, just to the highest frequency in the band-limited signal. In this way, the cosine Fourier series becomes a discrete cosine transform (DCT).

A picture is a two-dimensional object. The same principles that applied in the horizontal direction of the graph of Figure 3.6(a) can be applied in the vertical direction of a picture as well. Instead of the time variable in Figure 3.6, a distance variable is used. This is equally valid. (Of course, distance along a raster-scanned image has a relationship to time!) These principles also apply to both directions simultaneously. To apply the DCT, we need a periodic extension of the image. This is best done by breaking the image up into smaller blocks. The size of block chosen in most digital television applications is 8 picture elements (pixels) wide and 8 pixels high. This block can be imagined to repeat in a checkerboard pattern so that it is a symmetrical replication from checkerboard square to checkerboard square. This allows us to utilize only cosines and avoid the sines entirely. However, just as with the time waveform, this DCT will be valid only within the chosen block. Each block of the image will have to be handled separately. Again, bandwidth limiting of the video means that the image details are limited in the number of variations they can make within the block. This means that the DCT summations have a fixed maximum number and don't have to go to infinity.

The video samples in two directions are transformed into a form of spectrum by the two-dimensional DCT of Equations (3.6), (3.7), and (3.8)[13]:

$$F(u,v) = \frac{1}{4}C(u)C(v)\sum_{x=0}^{7}\sum_{y=0}^{7} f(x,y)\cos\left[\frac{(2x+1)u\pi}{16}\right]\cos\left[\frac{(2y+1)v\pi}{16}\right] \tag{3.6}$$

where

$$x \text{ and } y = \text{pixel indices within an 8-by-8 video block}$$

That is, x and y are the row and column numbers of the block that contains the measurements of the picture's brightness at each location in the block. Similarly,

u and v are DCT coefficient indices within an 8-by-8 spectrum block. That is, u and v are the row and column numbers of the block that contains the measurements of the picture's spectrum at each location in the block. Here,

$$C(u) = \frac{1}{\sqrt{2}} \quad \text{for } u = 0 \qquad \text{and} \qquad C(u) = 1 \quad \text{for all other values of } u \tag{3.7}$$

$$C(v) = \frac{1}{\sqrt{2}} \quad \text{for } v = 0 \qquad \text{and} \qquad C(v) = 1 \quad \text{for all other values of } v \tag{3.8}$$

To summarize, when we take a block 8 pixels wide and 8 pixels high from a bandwidth-limited video signal, there are 64 pixels. Each pixel is sampled in time. Its amplitude is measured, and the measurement is rounded to the nearest quantizing level. This quantizes the amplitude to a resolution of 8 bits (256 levels). We then have an array of 64 numbers in an 8-by-8 square (see Figure 3.7(a)).

If we apply Equations (3.6), (3.7), and (3.8) to this array, we will get another array of 64 numbers in 8 rows and 8 columns (see Figure 3.7(b)). This array represents the frequency components of the original amplitude array. The cells in this array contain numbers called the *frequency coefficients*. The number in the upper left-hand corner is related to the direct current (dc) component, that is, the average amplitude across the video square. The cells across the top row are the values of the horizontal frequencies in the block. The cells down the left side of the block are the vertical frequencies in the block. The other blocks represent frequencies in other directions (see Figure 3.7(c)).

The Fourier series works in both directions. If we have an amplitude waveform, we can convert it into a frequency series. If we have a frequency series, we can convert it back to the waveform. Similarly, there is an inverse discrete cosine transform (IDCT), which will convert the frequency coefficients array back into the amplitude array. Here are the equations for the IDCT[14]:

$$f(x,y) = \frac{1}{4}\sum_{u=0}^{7}\sum_{v=0}^{7} C(u)C(v)F(u,v)\cos\left[\frac{(2x+1)u\pi}{16}\right]\cos\left[\frac{(2y+1)v\pi}{16}\right] \tag{3.9}$$

$$C(x) = \frac{1}{\sqrt{2}} \quad \text{for } x = 0 \qquad \text{and} \qquad C(x) = 1 \quad \text{for all other values of } x \tag{3.10}$$

$$C(y) = \frac{1}{\sqrt{2}} \quad \text{for } y = 0 \qquad \text{and} \qquad C(y) = 1 \quad \text{for all other values of } y \tag{3.11}$$

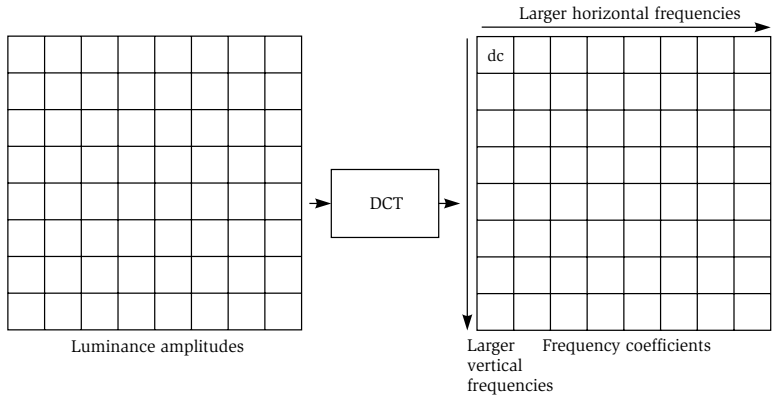
Applying these equations to the array of Figure 3.7(b) yields the array of Figure 3.8, which is recognized as the original amplitude array. This is a lossless conversion (to the limits of computational accuracy).

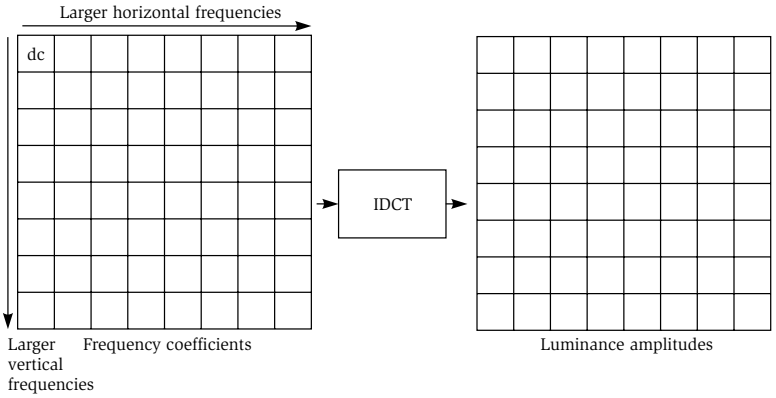
139	144	149	153	155	155	155	155
144	151	153	156	159	156	156	156
150	155	160	163	158	156	156	156
159	161	162	160	160	159	159	159
159	160	161	162	162	155	155	155
161	161	161	161	160	157	157	157
162	162	161	163	162	157	157	157
162	162	161	161	163	158	158	158

(a)

235.6	–1	–12.1	–5.2	2.1	–1.7	–2.7	1.3
–22.6	–17.5	–6.2	–3.2	–2.9	–0.1	0.4	–1.2
–10.9	–9.3	–1.6	1.5	0.2	–0.9	–0.6	–0.1
–7.1	–1.9	0.2	1.5	0.9	–0.1	0	0.3
–0.6	–0.8	1.5	1.6	–0.1	–0.7	0.6	1.3
1.8	–0.2	1.6	–0.3	–0.8	1.5	1	–1
–1.3	–0.4	–0.3	–1.5	–0.5	1.7	1.1	–0.8
–2.6	1.6	–3.8	–1.8	1.9	1.2	–0.6	–0.4

(b)





(c)

Figure 3.7 DCT 8-by-8 blocks. (a) Luminance values. (b) Frequency coefficients after DCT. (c) DCT and IDCT transformations.

139	144	149	153	155	155	155	155
144	151	153	156	159	156	156	156
150	155	160	163	158	156	156	156
159	161	162	160	160	159	159	159
159	160	161	162	162	155	155	155
161	161	161	161	160	157	157	157
162	162	161	163	162	157	157	157
162	162	161	161	163	158	158	158

Figure 3.8 Luminance values after IDCT.

The order of transmission of this information is shown in Figure 3.5. First, four blocks of luminance information are sent. Then one block of each of the two color difference signals is sent. The color difference signals come from sampling every other location in the luminance array of four blocks. All six of these blocks are separately converted by the DCT circuits and separately transmitted to the receiver.

Quantization

Let's review what has been accomplished so far: The video signal has been time sampled and quantized in a manner that gives the values of the luminance and color difference components at specific points in the picture. Those luminance points have been organized into blocks 8 pixels wide by 8 pixels high. The same has been done for the color difference signals, except every other pixel has been skipped to form macroblocks that span 16-by-16 luminance pixels. Then the values of these pixels have been converted to frequency values using the DCT in a manner that allows the original values to be computed again using the IDCT. This digitization and conversion by the DCT has not, by itself, done anything to reduce the redundancy of the image. More processing is needed to eliminate redundancy.[15]

The first step in the process is to note that there is a much wider range of values in the frequency domain cells than in the original video domain cells. Many of the frequency numbers are very small. In order to efficiently transmit the information in the frequency cells, we must reduce the amount of information necessary to convey it. It is already sampled and digitized, so the next step is to quantize the values in the frequency cells.

The transform coefficients represent different frequencies in brightness and color. Studies have shown that the human visual system has differing sensitivities to this information, depending on the amount of motion, the amount of detail, and the color of the image. There is a lot more information in this data than the human visual system requires for a satisfactory viewing experience. Consequently, the values in the cells can be approximated by a smaller set of values. The actual cell value is compared with the smaller set, and the nearest set value is chosen to represent the actual value. This process is another quantization mechanism. The cells are quantized differently according to the results of the research on human perception. A *quantizer matrix* is assigned to the transform coefficient matrix that prescribes the coarseness of quantization for each cell. Predetermined quantizer matrices are built into the system. Additionally, specific quantizer matrices can be downloaded as part of the signal for special cases.

Coarser quantization requires fewer bits, and the video matrix consisting of 64 cells with 8 bits in each row is transformed into 64 cells with fewer total bits. A most important result is that many of the high-frequency blocks contain coefficients that are so small that they become zero when quantized to a coarse scale. Thus the coefficient matrix is loaded with zeros. In addition, a quantizer coefficient is applied to the entire quantizer matrix to adjust its coarseness relative to other blocks in the picture. In this way, some areas of the picture are assigned more bits than other areas in relation to their importance to the image.

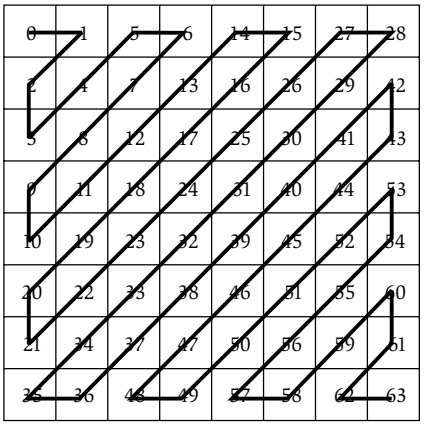
Some of the image's blocks will have little detail and many zeros in their coefficient matrices and thus will use few bits. Other blocks will have a great deal of detail and relatively few zeros in their coefficient matrices and will require many more bits. In this manner, bits are allocated to the parts of the picture that need them.

Run Length Encoding

The order of transmission of the values in the coefficient matrix is very important. A process called *zigzag scanning* is employed. Figure 3.9(a) illustrates this. In Figure 3.9(b), the frequency coefficients from Figure 3.7(b) are plotted as a graph in the order scanned in Figure 3.9(a). (In Figure 3.9(c), the large direct current (dc) term is not plotted so that the relative sizes of the remaining terms can be better seen.) The frequency coefficients quickly become small in value as the scanning progresses through the block. Most can be rounded to zeros, yielding a "run of zeros." Rather than transmit all the zeros, a pair of numbers is transmitted. The first number of the pair tells the number of zeros. The second number tells the value of the cell that terminates the run of zeros.[16] When a condition is reached where all the remaining cells have zeros, an end-of-block code is sent. Already, the number of bits to be transmitted has been significantly reduced.

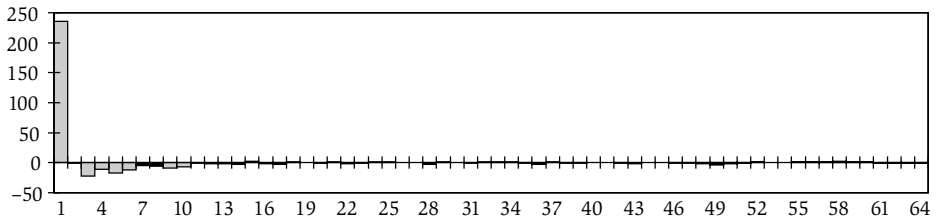
0	1	5	6	14	15	27	28
2	4	7	13	16	26	29	42
3	8	12	17	25	30	41	43
9	11	18	24	31	40	44	53
10	19	23	32	39	45	52	54
20	22	33	38	46	51	55	60
21	34	37	47	50	56	59	61
35	36	48	49	57	58	62	63

Luminance amplitudes

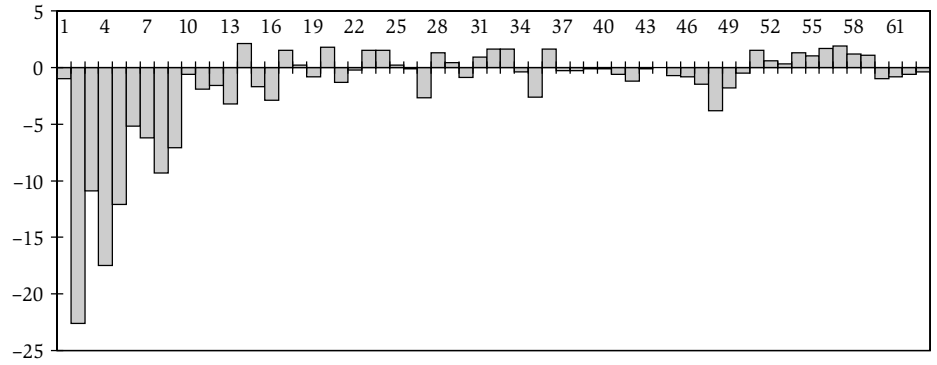


Zigzag scanning

(a)



(b)



(c)

Figure 3.9 DCT spectral block transmission order. (a) Zigzag scanning of the amplitude blocks. (b) Frequency coefficients scanned in zigzag order including dc component. (c) Frequency coefficients scanned in zigzag order without dc component.

Entropy Coding

The run length information occurs at a predetermined set of probabilities. Because of this, it is possible to employ entropy coding. As in Morse code, the most frequently occurring patterns are represented with the shortest codes, and the less frequently occurring patterns are assigned longer codes. This further reduces the number of bits needed to transmit the information. Huffman developed a method of creating such compression codes, called *Huffman codes*, when the probability of the message symbols is known. If another language is to be transmitted, another symbol table is used. If digitized video is to be transmitted, still another symbol table is applied.[17]

3.3.3 Coding of Moving Images

Now that we have completed the discussion of the compression of a still image (*intraframe coding*), we can extend the model to deal with moving images (*interframe coding*). Still-image compression ratios of 8:1 can typically be achieved with these techniques, but we still have a long way to go to reach a 50:1 ratio. Fortunately, all of the same principles we applied to a still image (transform coding, quantization, run length encoding, Huffman coding) can also be applied to the differences between one frame and the next. Two techniques are often used in conjunction in intraframe coding:

1. Calculate the difference between a macroblock in one frame and another (*picture differencing*).
2. Attempt to find the best match between a macroblock in one frame and that macroblock in another frame and transmit the motion vector (*motion compensation*).

Picture differencing breaks down if there is significant motion in the image. If you imagine a typical scene where the camera is panned across a scene, there are unlikely to be many similarities between a macroblock in one frame and that macroblock in the subsequent frame. The same is true for moving objects within the scene.

Motion compensation attempts to find where each macroblock has "moved" between frames, but it is unlikely that an exact match will be found, due to distortion caused by perspective and lighting changes. So when the best match is found, picture differencing is used to send the changes in the macroblock along with its new position.

Both of these techniques are used in conjunction in efficient video compression algorithms (such as MPEG).

Picture Differencing

In successive frames of video, each frame is typically very similar to the previous frame. If we take the limiting case of video of a still image, say, a landscape with

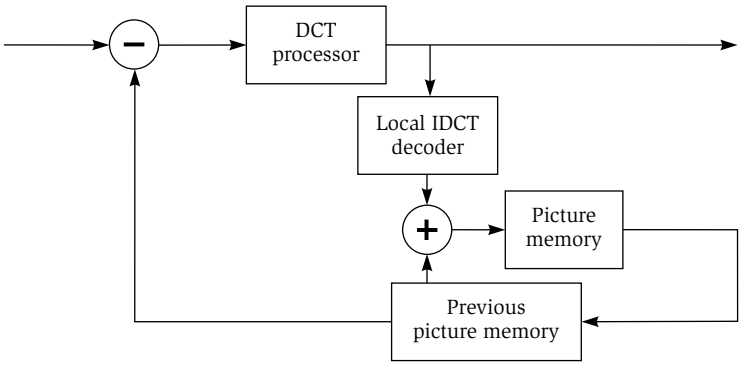


Figure 3.10 Interframe encoder.

no activity, there are no differences between frames. Sending one frame tells all. Succeeding frames simply can be repeated from the receiver's memory.

Figure 3.10 shows how the encoder can avoid sending the same frame repeatedly. When the equipment is turned on, all picture memories are empty. The first frame is sampled, and the DCT processor creates the frequency coefficient matrices for the luminance and color blocks of the picture and sends them to the IDCT decoder at the receiver site and to a local IDCT decoder. The local decoder produces the same image that will exist at the receiver. This image is then compared to the contents of the previous picture memory. Since there was no previous picture, the previous picture memory is empty, and there is nothing to add or subtract. The entire decoded picture is loaded into the picture memory. When the second frame is sampled, the previous picture is subtracted from it. Since they are identical, the subtractor produces zero values, the resulting coefficient matrix is all zeros, and the previous picture is unchanged. The process repeats itself, with no differences being transmitted to the receiver.

Figure 3.11 displays the decoder. Initially, its picture memories are empty. When the first frame's coefficient matrix arrives, it is decoded in the IDCT unit.

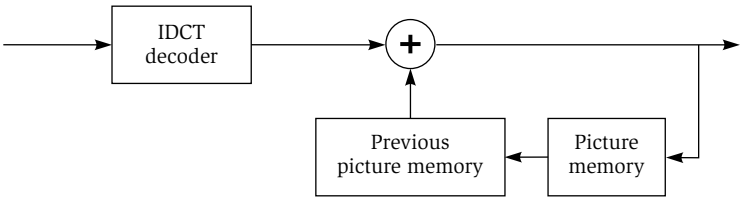


Figure 3.11 Interframe decoder.

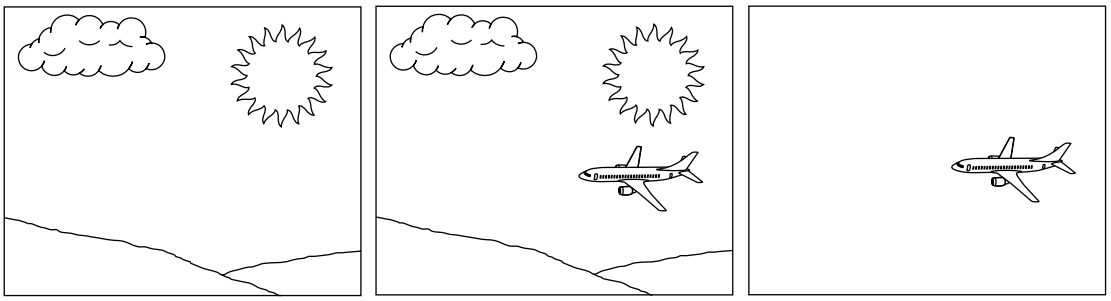


Figure 3.12 Differential transmission. (a) Original image. (b) Modified image. (c) Difference.

This is presented for display and for storage in the picture memory. At the time of the next frame, a zero coefficient matrix arrives. Nothing is added to the previous picture memory, and its contents are presented for display and are reloaded into the picture memory.

Watching a still picture on television, for example, the landscape image, shown in Figure 3.12(a), is not very interesting. Now let's assume that the picture changes because an airplane enters the picture (see Figure 3.12(b)). After the encoder subtracts the previous picture from the new picture, only macroblocks containing the image of the airplane have changed (see Figure 3.12(c)). These macroblocks may be encoded as differences from the previous picture, and a new coefficient matrix is transmitted to the receive site and the local decoder. (Note that the encoder may choose to intracode the macroblocks.) At the decoder, the coefficient matrix of the airplane is received and added to the previous landscape stored in the previous picture memory. It is presented for display and stored in the picture memory.

Two kinds of coefficient matrices have been transmitted in our example. The first coefficient matrix was a complete picture, called an intraframe (*I frame*), and all macroblocks are intracoded, or *nonpredicted*. The next frame contained information used to predict the contents of the picture memory for display; it is called a *P frame*. (Note that a P frame actually contains a mix of predicted macroblocks and nonpredicted macroblocks.)

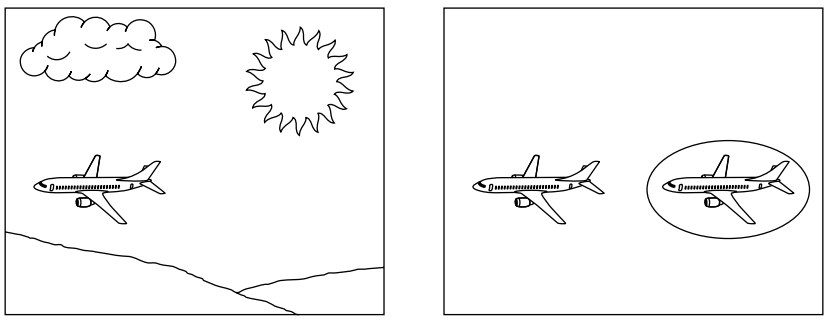
In a perfect world, we'd need just one I frame and all the rest could be P frames. There are a couple of reasons why this is insufficient. First, if the viewer missed the I frame while "surfing" channels, he or she would not get a correct image. In our example, the viewer might get just the airplane and not the landscape! A second problem is that there are imperfections and distortions that accumulate. It is important to reset the process every so often to limit the propagation of errors. So I frames are sent several times a second to allow channel surfers to have a starting point and to limit the propagation of errors.

Motion Compensation

Motion compensation can be employed to significantly reduce the redundancy even further. For example, in Figure 3.13(a) and (b), we note that the image of the airplane is the same in all appearances. It is redundant to have to retransmit the airplane; once should be sufficient! It should be adequate to simply convey an instruction to move the airplane to the new position. Of course, it is not that simple.

Motion compensation is based on the observation that much of the motion in images consists of a displacement of an element of the image. Motion compensation takes each macroblock and compares it with all possible macroblocks formed by making horizontal and/or vertical displacements in small increments (typically half-pixel). When the best fit is found, instructions are conveyed to the receiver to work with the contents of the macroblock and, when done, to relocate them to their new position with a *motion vector*. In our example, the encoder motion compensation would determine that the airplane has simply been moved horizontally. Instructions to the receiver would tell it to take the contents of a number of blocks in memory and move them to other locations that are some number of pixels displaced horizontally and vertically.

The motion compensation algorithm will usually find that the macroblock match is not perfect, and so it encodes any changes and transmits them with the motion vector. In our example, if the airplane made a slight rotation, simply moving it horizontally and vertically would not be enough. The difference between its previous pixels and its new pixels would be encoded. In most cases, this would be just a small fraction of the total pixels. Most would be unchanged. The resulting instructions to the receiver include taking blocks from memory, modifying them slightly, and then relocating them. When they are relocated,

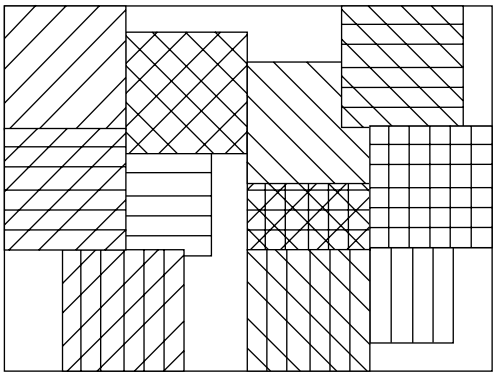


(a) (b)

Figure 3.13 More complex motion. (a) Further modified image. (b) Differences.

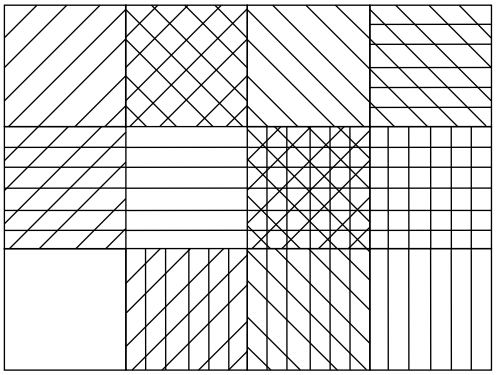
parts will overlay other blocks, and parts will be covered by other blocks. That information must be conveyed as well. Holes in the image will be created. They must be filled. The various parts of the image may move in different directions by different amounts. The motion compensation process can be very sophisticated to aggressively reduce the amount of information that needs to be transmitted (see Figure 3.14). Areas in the preceding frame of Figure 3.14 are found that best match the areas in the current frame. These are transmitted along with instructions for their new position and the differences needed to update them. This minimizes the amount of information that must be conveyed to the receiver.

Motion compensation makes the encoder much more complex than the decoder, which is acceptable because there are many more decoding



Preceding frame

Areas in preceding frame that most closely match corresponding areas in current frame



Current frame

Figure 3.14 Motion compensation.

sites than encoding sites. Since motion compensation is done at the encoder, viewers will enjoy improved results with their existing decoders as encoder motion compensation technology advances. The goal is to make receivers inexpensive.

B Frames and Groups of Pictures

Consider the airplane and landscape example shown in Figure 3.15. Five locations of the airplane occur. When the airplane moves from one position to another, it uncovers a piece of sky, which is present in subsequent images. If we had one of the subsequent images already in memory, for example, Figure 3.15(c), we could take the piece of sky needed from a "future" image. We could "predict backward" in time. The ability to predict both forward and backward in time is possible if we introduce a slight delay in the transmission and add memory to the receiver for more than one frame of video. The frames that allow bidirectional prediction are called *B frames*. Although they add cost to the decoder, they increase transmission efficiency. The trade-off in cost and complexity is governed by the cost of the memory. The cost of memory has come down substantially, so B frames are a worthwhile investment because bandwidth is the ultimate scarce resource.

We have defined three kinds of pictures: I frames, P frames (or predicted frames), and B frames (or bidirectionally predicted frames)*. They are clustered into a *group of pictures* (GOP) (see Figure 3.16). In Figure 3.16(a), the frames are presented in the order in which they are displayed. In order to minimize receiver complexity, they need to be transmitted in the order shown in Figure 3.16(b). In a sophisticated encoder, the GOP flow is interrupted at a scene change, and an I frame is transmitted (recall that an I frame contains all of the information needed to decode the frame within the frame itself and thus is not dependent on adjacent frames). However, I frames require much more data than other frames (a ratio of 5:3:1 for I:P:B frame size is typical). Therefore, the flow of data is very uneven, and the decoder must have a buffer of sufficient size to even out the flow.

Field-Based Prediction

All of the discussion so far has dealt with video as a sequence of frames. If the source material is generated by interlace scanning (as is the case with most video material), each frame is actually composed of two fields separated in time by the field period. For scenes with relatively little motion, the "double image" is

---

* More correctly, but less commonly, these are known as *I pictures, P pictures,* and *B pictures*. The *frame* terminology probably got started because MPEG was originally designed to code purely frame-based material and seems to have stuck.

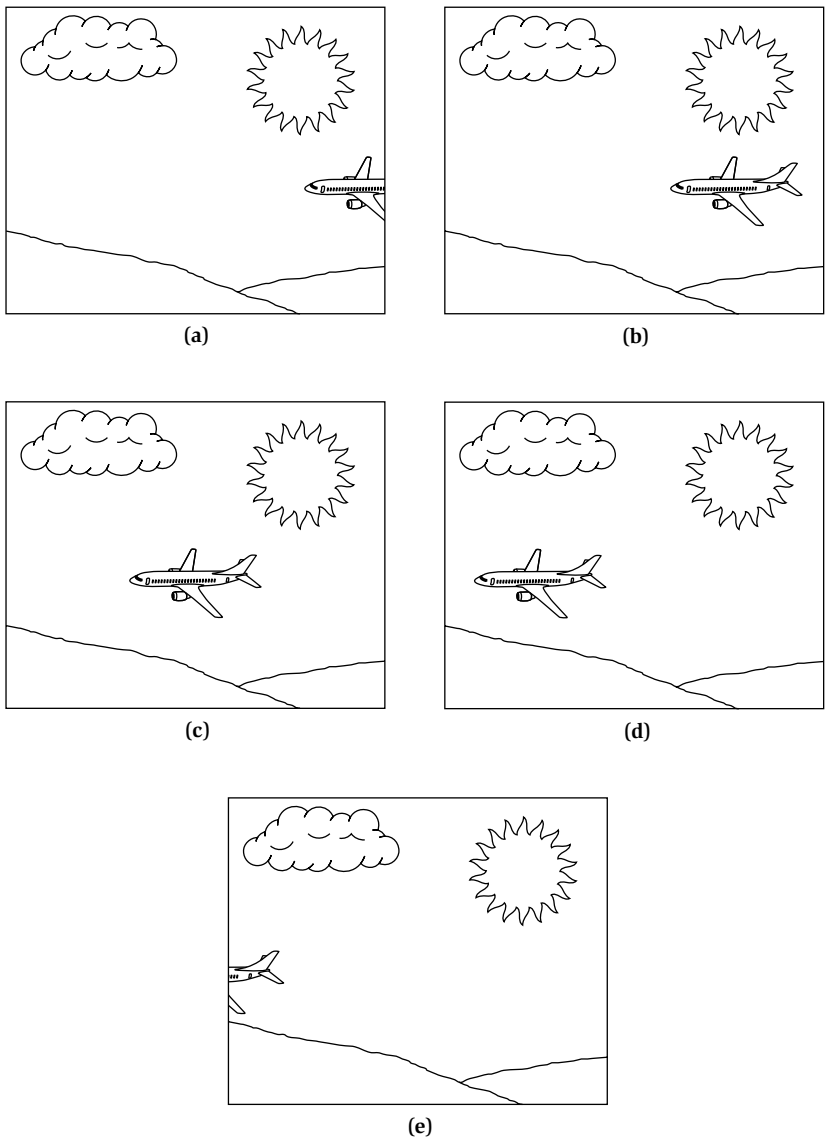


Figure 3.15 Motion compensation example.

not a problem and pure frame-based encoding may produce acceptable results. However, when there is rapid motion, the motion compensation algorithm needs to adaptively move into its field-based mode. The decision to code the field or the frame can be made at the macroblock level.

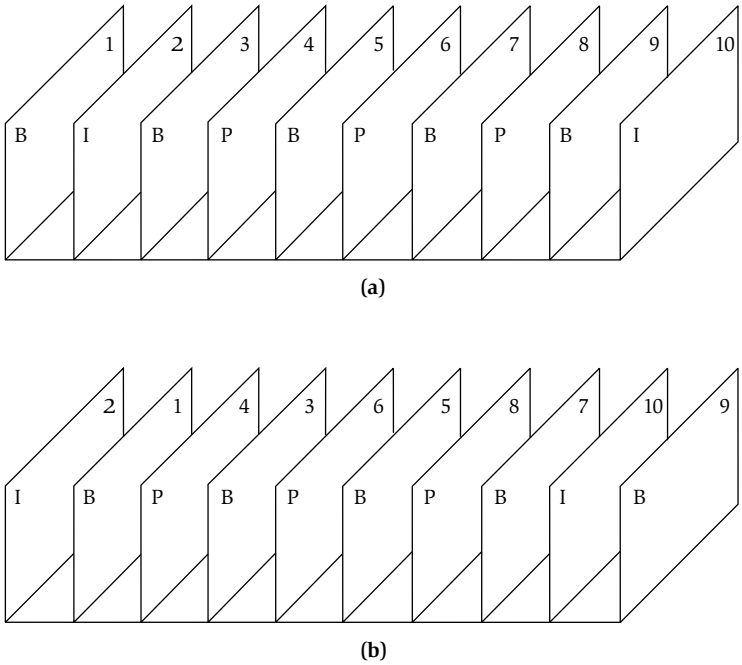


Figure 3.16 Order of transmission and display of I, P, and B frames. (a) Frame type and display order. (b) Frame type and transmission order.

If the field is coded, the values for the block are taken from alternate lines in the frame buffer — essentially reconstructing the field. After DCT, the coefficients for field 1 are sent, followed by those for field 2 for that macroblock. Picture differences and motion compensation for the block can reference either of the two fields in the preceding or succeeding frames.

Obviously, adaptive field/frame-based coding represents a considerable increase in complexity at the encoder and decoder, but it is essential to provide good results for interlaced material at moderate bit rates.

3.3.4 Digital Artifacts

Since each block of the image is treated separately, it is easy to appreciate that there may be differences in decoded video at the boundaries of the blocks. Techniques exist for smoothing over these boundaries. A digitally compressed image with inadequate data will occasionally display the boundaries of the blocks. This is called *blocking noise*.[18]

Recall that the 8-by-8 DCT and IDCT perform lossless conversion between the video and frequency space representations to the limits of computational precision. However, once the coefficients are quantized, the lossless nature of the computations is abandoned. Some distortion is introduced. If the distortion is minor, it might not be noticeable. But even if it is noticeable, it might not be objectionable; it may just blend in and get lost. However, under some circumstances, sharp edges will yield attenuated ripples. This is called the *Gibbs phenomenon*, after the mathematician who studied it.

If a sharp point exists in the image, the Gibbs phenomenon will exist in two dimensions. The resulting image distortion has been called *mosquito noise* or *the attack of the killer bees*. This artifact appears as a cluster of mosquitoes or killer bees surrounding the sharp point in the image. If a sharp edge exists in the image, ripples may precede and follow it.

3.3.5 Information Carried in the Vertical Blanking Interval

The astute reader will have noticed that only the active video lines in the NTSC signal are encoded. Clearly, there are several information fields normally carried in the vertical blanking interval (VBI), which would not make it to the television receiver:

1. Closed captioning, carried in line 21, field 1[19]
2. Nielsen source identification (SID) automated measurement of lineups (AMOL)
3. Teletext
4. VBI time code (VITC)
5. Vertical interval test signals (VITS)

Two SCTE standards[20,21] have been developed to encode these lines and insert them into the digital bitstream for carriage to the STT. Since the VBI information is considered inseparable from the video itself, it is carried in the video elementary stream by the picture user data field specifically designated for this purpose (see MPEG-2 Transport discussion later). After the video is decoded, the picture user data is used to reconstruct the VBI data, and it is inserted into the NTSC output signal by the NTSC encoder. In this way, backward compatibility is maintained with the installed base of analog televisions, which expect to see closed caption information in the VBI of the baseband analog signal.

Digital television has its own standard for closed captioning, specified by EIA-708,[22] which includes additional features above and beyond EIA-608, such as popup windows, extended characters sets, and the like. The FCC imposed a deadline of July 2002 to include EIA-708-compliant closed captions in digital television transmissions.

3.3.6 High-Definition Television

The discussion of digital video compression so far has focused on standard-definition television signals. The principles apply equally well to high-definition television, except that picture resolution and bit rates are increased considerably.

Sampling

HDTV has two main variants — a 720-line progressively scanned system (720p) and a 1080-line interlace scanned system (1080i). The latter is based on the SMPTE 240M standard for 1125/60-line HDTV production systems.[23] Sampling parameters for the two systems is shown in Table 3.2 (standard-definition television is included for comparison).

A 1080i picture has six times as many active pixels as a standard-definition picture, and hence the raw information rate is increased by a factor of 6. The development of high-definition encoders has provided a significant challenge to chip manufacturers, a challenge that has been met by a combination of parallel processing and higher clock rates, made possible by the reduction in integrated circuit line widths.

Block formation

Although the picture resolution is higher, the block size remains the same; there are just more blocks per picture. In addition, macroblocks are formed in the same way. Standard-definition television uses the MPEG-2 main profile at main level (MP@ML), while high-definition television uses the MPEG-2 main profile at high level (MP@HL), (see Section 3.3.7). The SCTE has defined the allowed set of resolutions and frame rates that are supported by cable systems as shown in Table 3.3.

3.3.7 MPEG Standards

The Motion Picture Experts Group (MPEG) of the Geneva-based International Standards Organization (ISO) has been active in video compression standards since the early 1990s. All MPEG standards are based on a set of principles that have allowed tremendous innovation while maintaining an interoperable standard.

Table 3.2 Standardized Video Input Formats (from [26], used with permission)

Video standard	*Active lines*	*Active samples/line*
ITU-R BT.601-5	483	720
SMPTE 296M[24]	720	1,280
SMPTE 274M[25]	1,080	1,920

Table 3.3 Compression Format Constraints for Digital Video (from SCTE 43 2001[26] used with permission)

Vertical resolution	*Horizontal resolution*	*Aspect ratio*[a]	*Frame rate*[b]	*Progressive sequence*[c]
1,080	1,920	1, 3	1, 2, 4, 5	1
1,080	1,920	1, 3	4, 5	0
1,080	1,440	3	1, 2, 4, 5	1
1,080	1,440	3	4, 5	0
720	1,280	1, 3	1, 2, 4, 5, 7, 8	1
480	720	2, 3	1, 2, 4, 5, 7, 8	1
480	720	2, 3	4, 5	0
480	704	2, 3	1, 2, 4, 5, 7, 8	1
480	704	2, 3	4, 5	0
480	640	1, 2	1, 2, 4, 5, 7, 8	1
480	640	1, 2	4, 5	0
480	544	2	1	1
480	544	2	4	0
480	528	2	1	1
480	528	2	4	0
480	352	2	1	1
480	352	2	4	0

MPEG-2 coded values:
[a] 1 = square pixels; 2 = 4:3; 3 = 16:9.
[b] 1 = 23.976 Hz; 2 = 24 Hz; 3 = 25 Hz; 4 = 29.97 Hz; 5 = 30 Hz; 6 = 50 Hz; 7 = 59.94 Hz; 8 = 60 Hz.
[c] 0 = interlaced scan; 1 = progressive scan.

- MPEG standards do not specify the operation of the encoder, only that of a reference decoder and the syntax of the bitstream that it expects. This allows encoder implementations to improve over time, for example, as better motion compensation search algorithms are developed.
- MPEG standards are a toolbox approach — various features are added from one "profile" to another, each higher profile adding complexity and efficiency.

MPEG-1

MPEG-1,[27] which was standardized in November 1992, introduced efficient coding for frame-based (progressively scanned) material based on DCT coding of 8 by 8 blocks. MPEG-1 is based on the intraframe coding principles from JPEG,

but it adds interframe coding techniques, including motion compensation and the associated difference pictures, the so-called P and B frames. MPEG-1 supports only low resolutions (352 × 240), which is insufficient for entertainment-quality video. (However, it was quickly extended to higher resolutions, for example, the Full Service Network[28] used so-called "MPEG-1.5" at a resolution of 352 × 480. DirecTV also launched with MPEG-1 at higher resolution, subsequently migrating to MPEG-2 before the transport packet structure was finalized.) Although MPEG-1 works well for progressively scanned material (for example, film), it requires high bit rates (6 Mbps or greater) to compress NTSC video. This is primarily because MPEG-1 does not support field-based coding.

MPEG-2

MPEG-2 was standardized in November 1994. MPEG-2 adds support for adaptive field/frame coding, which is particularly for encoding interlaced material with high motion content (for example, sports). (MPEG-3 was originally targeted at HDTV, but it was abandoned when it became apparent that MPEG-2 was easily extendable to support HDTV.)

MPEG-2 has been a tremendously successful standard. It has been adopted by the television and cable industries worldwide and also by DVD. MPEG-2 provides a toolbox comprising a collection of image-processing components; a specific application uses appropriate portions of the standard. Table 3.4 summarizes the profiles and levels defined by the MPEG-2 standard (some levels and profiles have been omitted for clarity). The "level" describes the amount of video resolution. The "profile" describes the principal features of that collection of components. The general-purpose profile is called the *main* profile. The *simple* profile drops the B frames and simplifies the decoder by reducing the

Table 3.4 MPEG-2 Levels and Profiles (Subset)

		Profile			
		Simple	*Main*	*High*	*4:2:2 (studio)*
	Pictures	*I, P*	*I, P, B*	*I, P, B*	*I, P, B*
Level	*Sampling*	*4:2:0*	*4:2:0*	*4:2:0 or 4:2:2*	*4:2:2*
High	Resolution		1920 × 1152	1920 × 1152	
	Frame rate		60 fps	60 fps	
Main	Resolution	720 × 480	720 × 480	720 × 480	720 × 512
	Frame rate	29.97 fps	29.97 fps	29.97 fps	29.97 fps
	Resolution	720 × 576	720 × 576	720 × 576	720 × 608
	Frame rate	25 fps	25 fps	25 fps	25 fps
Low	Resolution		352 × 288		
	Frame rate		30 fps		

memory and processing required. The trade-off is that the bit rate generally goes up. Standard-definition television uses the main profile at main level (MP@ML), while HDTV uses the main profile at *high* level (MP@HL), although there is a separate high profile designed with HDTV in mind. In addition, a 4:2:2 profile is defined for studio use.

MPEG-4

MPEG-4 was originally focused on "very-low-bit-rate coding" (as described in the 1993 RFI), but was changed the following year to "coding of audiovisual objects." MPEG-4 is very different in focus from previous MPEG standards, in that it focuses on multimedia and interactivity. Visual scenes are broken down into objects and sent as separate layers, which are composed at the decoder.

MPEG-4 also introduces some innovative techniques for very-low-bit-rate coding. These include mapping of images onto a computer-generated mesh — it is possible to map facial expressions and send them independently from a person's image and then reconstruct the animated face at the receiver.

Most of MPEG-4 is not intended for entertainment-quality video and is not very relevant to the cable industry. However, MPEG-4 Part 10, Advanced Video Coding, is a new standard for entertainment-quality video that offers significantly lower bit rates than MPEG-2.

MPEG-4 Part 10, Advanced Video Coding (AVC)

MPEG-4 Part 10 was completed in March 2003. It was developed by the Joint Video Team (JVT), a joint effort of the ITU-T Video Coding Experts Group (VCEG) and MPEG. It is also known as H.26L.

MPEG-4 AVC adds improvements to DC and AC prediction, sprite coding, error resiliency coding, nonlinear dc quantization, and has more accurate motion compensation (quarter-pel over MPEG-2's half-pel). MPEG-4 AVC also uses spatial prediction for intraframe coding to reduce the size of I frames. The basic spatial transform is an integer transform (an approximation of the DCT) that operates on a smaller, 4×4 pixel block. To improve motion compensation efficiency, MPEG-4 AVC adaptively uses different block sizes (4×4, 8×4, 4×8, 8×8, 16×8, 8×16, or 16×16) and multiple reference frames. Finally, postfiltering of the image is much more sophisticated, to help reduce blocking artifacts, and includes a deblocking filter in the prediction loop.

Just how much better is MPEG-4 AVC than MPEG-2? It is still too early to say at the time of this writing, but substantial improvements in coding efficiency are already being demonstrated by MPEG-4 AVC implementations. As with MPEG-2, we expect to see an improvement in coding efficiency over time. For example, from the MPEG-2 curve shown in Figure 3.17, it is apparent that MPEG-2 has almost reached the limitations of the fundamental principles upon which it is based. Bit rates much below 2 Mbps for entertainment-quality video do not seem

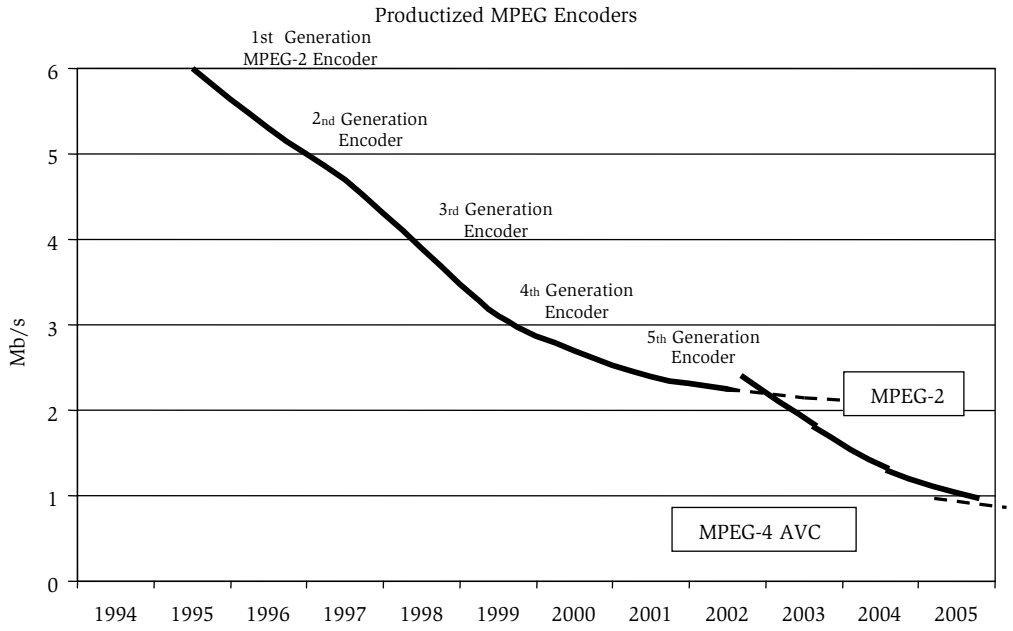


Figure 3.17 MPEG-4 AVC versus MPEG-2 bit-rate. (Courtesy of Harmonic Inc.)

likely. In contrast, MPEG-4 AVC shows the potential to be at least twice as efficient as MPEG-2 for the same perceptual quality.

However, there are significant issues of backward compatibility that prevent cable operators from adopting of MPEG-4 AVC, attractive as it may be from the point of view of bandwidth efficiency. SCTE digital television standards are based on MPEG-2 video compression. Moreover, a migration to MPEG-4 AVC would necessitate the wholesale replacement of tens of millions of decoders. Since video on demand is not currently subject to SCTE standards, it could employ MPEG-4 AVC to achieve more VOD streams per QAM channel, but at the cost of a new STT per VOD household.

MPEG-4 AVC is being seriously considered for high-definition DVD9 by the DVD forum. DVD9 is a single-sided, dual-layer DVD format with a total capacity of 8.5 GB. Toshiba demonstrated storage of approximately 135 minutes of high-definition video using MPEG-4 AVC encoding at 7 Mbps at the 2003 Consumer Electronics Show.

The issues of backward compatibility do not affect MPEG-4 AVC adoption in the Internet environment. Several software decoders were already available at the time of writing, and Apple has adopted MPEG-4 AVC as the core for QuickTime. Internet streaming media will probably enjoy a considerable advantage in

coding efficiency over broadcast television services (cable, satellite, and terrestrial broadcast).

3.3.8 Other Approaches to Video Compression

A variety of other approaches to compression exist besides the discrete cosine transform. The DCT has enjoyed the most attention because it is based on the Fourier transform, which has been studied more than any other transform. The Fourier transform has other broad applicability to the study of spectra and signal analysis. It has received extensive funding in research on radar, sonar, and deep-space signal analysis. Efficient computational methods have been developed to apply it. Custom-designed integrated circuits have been constructed to implement very fast conversion of waveforms to spectra, and vice versa. The DCT is a natural starting point for video compression.

Other approaches will continue to be studied and may at some point close the gap that currently exists between DCT-based and other compression techniques. Efficient software and hardware may develop for these other approaches, and eventually the DCT may be displaced. We will summarize three alternatives — wavelets, fractals, and vector quantization — that have all been around for a considerable time but have not seriously challenged DCT-based approaches.

There will continue to be further advances in video compression (like MPEG-4 AVC). It is therefore important to avoid building regulatory or other mechanisms that will stand in the way of implementing advancing technology. The recent history of the personal computer industry has shown that the marketplace will tolerate some degree of difficulty with interoperability and compatibility if the technical advances are substantial. This comes at a heavy toll in consumer frustration and sometimes anger. However, the old model of rigid compatibility may fall to the rapid advance of technology. In the end, the marketplace decides.

Wavelet

Wavelet transforms appear to be more “scalable” than other techniques. That is, they have the ability to serve a low-resolution display and then, with little waste of what has been transmitted, add more and more detail. A low-resolution display is able to use simpler circuits to extract what is needed, whereas a higher-resolution display can support the cost of more complex circuits and extract more information from the signal. A special class of filters is used to implement wavelets. As more research is done, more cost-effective implementations will become available.

Fractal

Briefly, fractals are equations that describe images. They may be the most efficient compression method, representing the most complex images with the least data. The challenge with fractals has been in finding the equations.

Vector Quantization

Vector quantization involves approximating the image out of a prestored table of image elements (sometimes called *stamps*). This "library" is used to build up the image. Choosing the elements to place in the library is critical to the success of vector quantization. Simpler images place lower demands on the library; images that are more complex are more challenging.

3.4 Digital Audio Compression

The human auditory system is considerably less tolerant of imperfections than the human visual system. Compression of quality audio is much more difficult than the compression of video. Compression ratios of only 4:1 or 8:1 are achieved in audio, whereas video enjoys compression ratios of 30:1 or 50:1. Fortunately, the audio bandwidths are small, and the resulting data rates are not too burdensome.[29]

Two major competitive methods of audio compression are in widespread use in digital television systems: MPEG-1 Layer 2 and Dolby AC-3. Much of Europe and the U.S. DSS satellite system have chosen MPEG-1 Layer 2, whereas the U.S. cable industry and the U.S. broadcast standards use AC-3. DVD uses Dolby AC-3 for 60-Hz systems (mostly U.S. and Japanese). Some DVDs (digital video discs) use MPEG-2 audio, but only the backward-compatible mode to MPEG-1 Layer 2, for 50-Hz scanning systems (mostly European).

A third method, MPEG-1 Layer 3, has become extremely popular for audio-only applications, and is generally known by its Windows file-extension name, "MP3."

3.4.1 Principles of Audio Compression

As with video, compression is achieved by the quantization of the frequency components of the signal.

The ear is extraordinarily sensitive to certain impairments in the audio signal, and spectral flatness is very important. Nevertheless, the ear is most sensitive to frequencies between 2 kHz and 5 kHz, and sensitivity drops quite rapidly above and below these frequencies. In fact, at 50 Hz and 17 kHz the sensitivity of the ear is about 50 dB lower than at 3 kHz.

Another effect of the psychoacoustic system is audio masking. A louder sound will mask a quieter one, especially if they are close in frequency. (This effect causes the white noise of the air conditioning fan in a hotel room to mask sounds from outside the room as long as they are not too loud.) The phenomenon is quite complex and very nonlinear. It is more effective when the quieter sound to be suppressed is higher in frequency than the louder sound. The extent to which a louder sound masks a quieter sound is called the *psychoacoustic masking threshold*. Perceptual coding takes advantage of this masking effect.

Figure 3.18[30] shows a high-level block diagram of an audio encoder–decoder pair. In the encoder, an analysis filter bank divides the input signal into narrow frequency bands, each of which may then be quantized and coded independently. The perceptual model adjusts the quantization level of each band such that quantization noise is ideally just below the masking threshold. The resultant bitstream is encoded and transmitted over a data link to the decoder. At the decoder, the inverse functions are applied to reconstruct the audio signal for playback to the listener.

Here are the main steps in the audio compression process.

1. Sampling and preprocessing
2. Sub-band coding
3. Transform coding (optional)
4. Quantization according to psychoacoustic masking threshold
5. Multichannel bit allocation (optional)
6. Huffman coding

Entertainment-quality audio is generally sampled at 48 kHz, with quantization to 16, 18, 20, or 24 bits per sample. (Compact discs are the exception, with 44.1 kHz-sampling at 16-bit quantization.)

Sub-band coding divides the audio spectrum into narrow frequency bands and codes each sub-band separately. The advantage of this approach is the quantization of each sub-band can be set independently according to (a) the sensitivity of the ear at that frequency and (b) the masking effect of signals in adjacent sub-bands. More advanced audio coding systems may use transform coding in conjunction with sub-band coding or to completely replace sub-band coding.

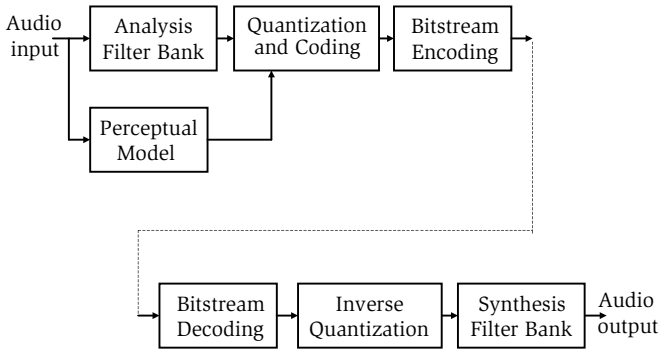


Figure 3.18 Audio Encoding and Decoding.

As each channel is coded, bits are dynamically allocated between sub-bands to achieve the best fidelity for a given bit rate. Huffman coding is used to code frequently quantized values in the same way as for DCT coefficients in video coding.

For multichannel audio, there is often redundancy between channels that can be exploited by multichannel bit allocation schemes. In particular, the ear locates sound sources better at some frequencies; outside of these frequency ranges two channels can be "bridged" together.

There are many excellent texts that discuss audio compression in more detail.[31,32]

3.4.2 AC-3 (Dolby Digital)

The AC-3 system has 5.1 channels of audio: front left, front right, center, surround left, surround right, and low-frequency enhancement (LFE). The bandwidth of the channels is around 20 kHz, with the exception of the LFE channel, which is limited to 120 Hz. This gives rise to the ".1" of the 5.1 channels. Since low-frequency sounds are not localized, only one speaker is needed for the low frequencies, and it can carry the low frequencies of all the channels. Additionally, the LFE speaker (popularly known as a "subwoofer") can be placed almost anywhere in the room.

Audio is compressed by first converting it to digital form with a sampling rate of 48 kHz. The audio is then segmented into *frames* of samples 1,536 samples long. The analog audio is quantized to at least 16-bit resolution. The frame is divided into six 512-sample *blocks*. Since the human auditory system is so sensitive to imperfections, half of the samples in each block overlap half of the samples from the adjacent blocks. That is, there is a 25% overlap with the preceding block and a 25% overlap with the following block. The sensitivity of the human auditory system to discontinuities requires doubling the number of transmitted samples! The samples are *windowed*; that is, the set of samples is multiplied by a weighting function that tapers the amplitude of the samples at the beginning and at the end of the sample set. The receiver blends the ending tail from one set with the beginning tail from the next set to avoid sudden changes or clicks or pops.

The audio samples are transformed into the frequency domain as a set of spectral data points using a modified version of the discrete cosine transform called the *time division aliasing cancellation* (TDAC) transform. Its special properties allow recovery from the need to have block overlap that would otherwise generate twice as many frequency coefficients. The result is that the 512 samples of an audio block yield only half as many, 256, frequency coefficients. These frequency coefficients divide the audio spectrum into 256 subbands of less than 100 Hz each. Up to this point, the processing of the digital samples is fully reversible, and only digitization has been accomplished. Compression has not yet been achieved.

Data compression is achieved by quantization and entropy coding. The psychoacoustic model dictates the degree of coarseness that can be accepted in the quantization of the audio spectrum data. The frequency coefficients are transmitted in the mathematical format known as *scientific notation*. A basic number is multiplied by an exponent. The basic number is called the *mantissa*. The same exponent generally applies to several adjacent blocks of samples, and data transmission can be reduced by taking advantage of this. Only changes are conveyed. Extensive techniques exist for reducing the amount of data that must be transmitted. The optimum bit rate for 5.1-channel Dolby AC-3 is 384 kb/s, although other rates are possible, yielding varying qualities of the resulting decoded audio signals.

AC-3 provides two important features to manage audio level and dynamic range at playback.

Dialog normalization: A subjective loudness of the dialog is explicitly coded in the AC-3 bitstream to allow balance to be achieved from one program segment to another. This is necessary because the level of the dialog may vary greatly between, for example, a movie and a commercial.

Dynamic range description: The dynamic range of the playback may also be adjusted at playback. This allows a listener with a home theater to enjoy the full dynamic range of a movie while also allowing the same movie to be enjoyed on a television with a small speaker.

Finally, AC-3 also defines an audio frame structure, which includes a method for synchronization and CRC-16 error detection codes. If an error is detected in the bitstream, the decoder will mute the audio and attempt to resynchronize. AC-3 was selected as the audio standard by the Advanced Television Systems Committee (ATSC) Digital Television Standard.[33] It is also specified by the standards for digital cable systems of the Society of Telecommunications Engineers (SCTE).

3.4.3 MPEG Audio

MPEG-1 Audio

MPEG-1 audio[34] coding defines three *layers*, which are the audio equivalent of video profiles. Each layer adds complexity but reduces bit rate. MPEG-1 audio was originally based on principles from the MUSICAM (masking-pattern adapted universal sub-band integrated coding and multiplexing) system.

MPEG-1 layer 2 is commonly used in satellite applications and in European broadcast applications. It uses 32 sub-bands (using quadrature mirror filtering) and codes entertainment-quality music at 128 Kbps per channel.

MPEG-1 layer 3 further subdivides the 32 sub-bands of layer 2 into 576 sub-bands by using a frequency transform (modified discrete cosine transform). It can provide about the same quality as MPEG-1 layer 2 using only 64 Kbps per

channel. MPEG-1 layer 3 is better known as "MP3" (from the file extension .mp3) and has become the de facto audio encoding algorithm used for "ripping" CDs and sharing them on the Internet. MP3 is also widely used in portable music players.

MPEG-2 Audio

MPEG-2 audio has two main variants: backward-compatible audio[35] and advanced audio coding (AAC).[36] Backward-compatible audio adds support for 5.1 multichannel audio while maintaining bitstream compatibility with MPEG-1. AAC is a second-generation algorithm that provides a considerable increase in coding efficiency over MPEG-1 audio. Good results are obtained at 64 Kbps per channel. AAC is also very flexible, providing up to 48 channels of audio, 16 low-frequency effects channels, 16 overdub/multilingual channels, and 16 data streams.

3.5 Digital Audio–Video Transport

So far, the discussion has covered how we take individual video feeds and audio feeds and convert them into "elementary streams," which are their compressed digital bitstream equivalents. The next step is to combine these elementary streams into a *transport stream* with some kind of structure so that an STT can select and decode the digital bitstreams according to our familiar program paradigm for broadcast television.

Time division multiplexing (TDM) allows multiple lower-rate bitstreams to be combined to form a single higher-rate bitstream. However, familiar TDM methods used for telephony (for example, SONET) are designed to carry many channels that are all exactly the same rate. These would not be suitable for video and audio elementary streams with arbitrary, and possibly variable, transmission rates.

An approach more like asynchronous transfer mode (ATM) is required to allow each elementary stream to be transmitted at the precise rate it requires. ATM approaches divide the bitstream into fixed-length *packets* (also known as *cells*) and allocate packets between channels based on their required bit rate. A wider channel is created by allocating more packets per second than a narrower channel.

The MPEG-2 Systems Standard[37] defines an ATM-like method to transport digital video, audio, and auxiliary data in a multiple program transport stream (MPTS). It is the basis for the ATSC Digital Television Standard and is specified by the SCTE Standards for Digital Television.

3.5.1 MPEG-2 Systems Layer

The MPEG-2 systems layer defines the two different ways to multiplex elementary streams from audio and video encoders, as illustrated in Figure 3.19:

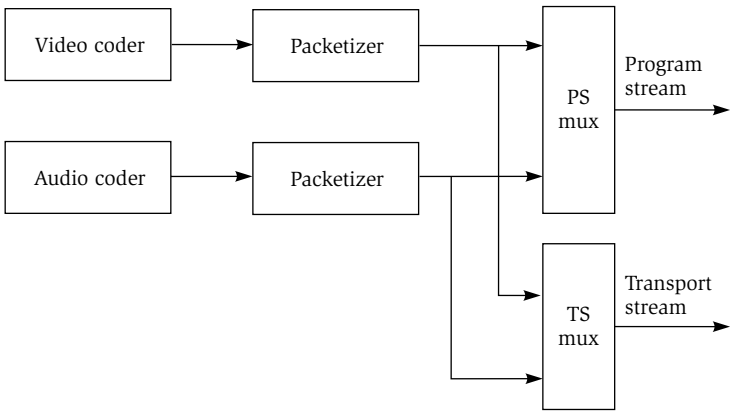


Figure 3.19 Coding, packetization, and multiplexing.

Program stream — designed for storage-based applications, for example DVD.[38] The program stream format is not generally used in cable television and will not be discussed further.

Transport stream — designed for transmission over a constant-delay network, for example, terrestrial broadcast and cable systems. The remainder of this discussion will focus on the transport stream.

Before the individual video or audio elementary bitstreams are combined to form a transport stream, they must have timing information inserted to describe the rate at which they should be displayed at the decoder, and they must be packetized so that they can be recovered at the receiver.

Figure 3.20 shows the structure of the packetized elementary stream (PES) packet. The start code prefix identifies the start of the packet. The stream ID identifies the type of stream (16 for video, 32 for audio). The packet is of variable length, as described by the PES packet length. A great deal of information is provided in the PES header (signaled by PES header flags), most importantly the presentation time stamp (PTS) and decode time stamp (DTS), which are used by

3 bytes	1 byte	2 bytes	2 bits	14 bits	1 byte	Variable length	
Packet start code prefix	Stream ID	PES packet length	10	PES header flags	PES header length	PES header fields	PES packet data block

Figure 3.20 Packetized Elementary Stream (PES) structure.

the receiver to synchronize the playout of audio and video. Since the PES packet is intended for error-free environments, there are no fields for error detection or protection.

The MPEG-2 systems transport stream supports an important set of features for broadcast digital television systems:

1. Multiplexing
2. Navigation
3. Timing distribution
4. Conditional access
5. Splicing

Multiplexing

Multiplexing allows many packetized elementary streams to be combined into a *multiple-program transport stream* (MPTS). This has numerous advantages for the operator. It allows the digital channel to be packed with multiple programs (analogous to television channels) so that the capacity of the channel is used efficiently. Additional auxiliary data can also be combined into the multiplex to carry program guide information, conditional access control messages (see Chapter 21), emergency alert information, and so on.

Figure 3.21 displays how packetized elementary streams are time division multiplexed to form a transport stream. Transport streams are identified by a unique packet identifier (PID) so that they can easily be filtered and routed to the different circuits and subsystems in the receiver that process and decode those data streams into the output signals that drive the presentation devices. Video and audio data must be intermixed in a manner that ensures that sufficient audio and video data arrive for their respective processors so that the sound and picture can be reproduced in synchronism. This figure is generic, and the concept applies to all transmission paths. The transport stream can be a

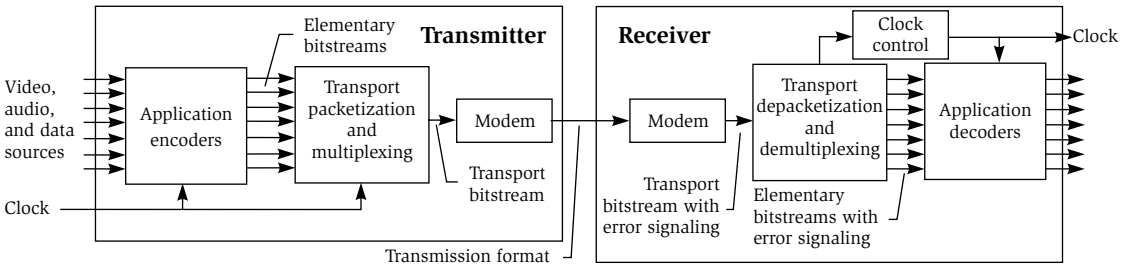


Figure 3.21 Transport packetization and multiplexing.

combination of a variety of packetized elementary streams. For example, the same video PES can be associated with several audio PESs in multiple languages. Likewise, the same audio PES can be associated with several video PESs, providing multiple views of the same athletic event. This packetization is necessary if there is to be local insertion of programming and commercials. This, of course, is a part of the current business structure in analog systems and is a critical part of the economics of the business.

Figure 3.22 shows the MPEG-2 transport packet format. Packets are fixed-length (188 bytes) with a fixed-length header (4 bytes). A sync byte having a fixed value of 0x47 identifies the start of the packet. The header contains packet identification, scrambling, and control information. A continuity counter increments for each packet and allows a missing or out-of-sequence packet to be detected.

The header is followed by 184 bytes of MPEG-2 or auxiliary data. Optionally, adaptation headers may be inserted to carry specific information that is not required in every MPEG-2 packet, for example, the program clock reference (see upcoming subsection on Timing Distribution). The Packet Identifier (PID) has local significance within the system multiplex: It makes it possible to demultiplex the transport stream at the receiver and to select the program of interest. Alternatively, one program can be processed for display while another one is recorded for later use. The programs recorded do not need to be further processed. That can take place at playback.

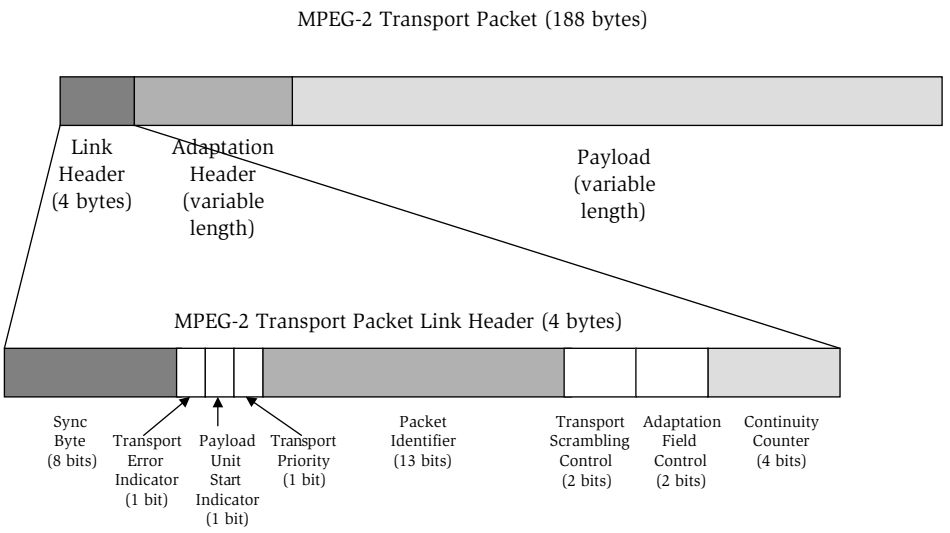

MPEG-2 Transport Packet (188 bytes)
Link Header (4 bytes)
Adaptation Header (variable length)
Payload (variable length)
MPEG-2 Transport Packet Link Header (4 bytes)
Sync Byte (8 bits)
Transport Error Indicator (1 bit)
Payload Unit Start Indicator (1 bit)
Transport Priority (1 bit)
Packet Identifier (13 bits)
Transport Scrambling Control (2 bits)
Adaptation Field Control (2 bits)
Continuity Counter (4 bits)

Figure 3.22 MPEG-2 Transport Packet Format. (From [39] used with permission)

The system multiplex can be extremely complex. Individual packetized elementary streams can be multiplexed together to form a program that includes multiple audio streams, multiple video streams, and ancillary signals such as conditional access control messages. This is called a *single-program transport stream* (SPTS). If the data capacity of the channel is sufficient to carry more than one SPTS, they can be multiplexed into a multiple-program transport stream (MPTS). In this way, many programs can be time division multiplexed into a transport stream of an arbitrary bit rate. The entire transport stream is then modulated onto a carrier signal to form a "digital" channel. Many channels can be combined to form a frequency division multiplex of several hundred megahertz.

Figure 3.23 shows an alternative view of a system multiplex. The bottom layer is the forward application transport (FAT) channel, which contains the MPTS. It is divided into PES packets, carrying compressed audio and video bitstreams, and program-specific information (PSI), carrying tables that describe the system multiplex. The next subsection, on Navigation, will describe how these are used by the receiver to find and select the desired program.

Navigation

Although the MPEG-2 transport stream syntax includes some basic structures to enable the receiver to select one out of many programs included in an MPTS, these structures are not sufficient to support the navigation features demanded by a sophisticated receiver. Moreover, they do not provide information to support an electronic program guide (EPG). We will describe the structures defined by the MPEG systems layer before moving on to necessary extensions introduced by the SCTE and the ATSC.

The MPEG-2 system multiplex always includes a program association table (PAT), which is transmitted continuously on PID value 0. The PAT contains an entry for each program within the multiplex and information identifying the PID

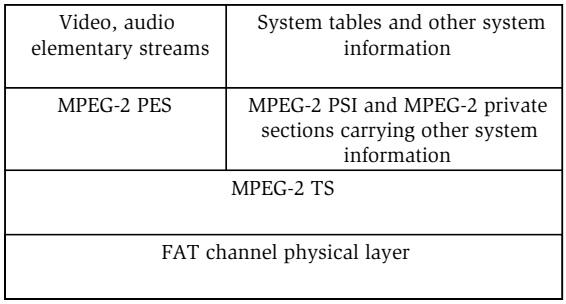

Video, audio elementary streams
System tables and other system information
MPEG-2 PES
MPEG-2 PSI and MPEG-2 private sections carrying other system information
MPEG-2 TS
FAT channel physical layer

Figure 3.23 FAT Channel Transport Layer Protocol. (From SCTE 40 2001)

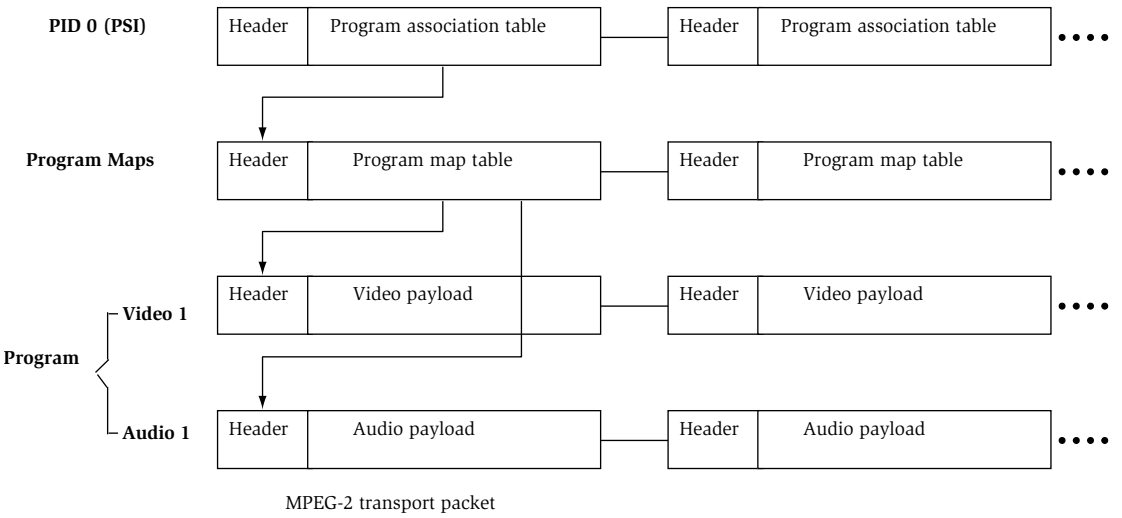


Figure 3.24 Program Association and Program Map Tables. (From [40] used with permission)

of the program map table (PMT) for it. In turn, the PMT contains the video PID(s), audio PID(s), and associated data PID(s) for the program.

Figure 3.24 illustrates how a receiver "navigates" the PAT and PMT to discover the video and audio PESs that make up the program. However, this is a very basic form of navigation; there is no description of the program name (for example, "PBS") or the program description (for example, "World News Hour").

In cable systems, information about the frequency and description of each channel is typically sent in a separate out-of-band data channel. This information is called the *channel map*. When digital transmission was introduced, it was a simple extension to add information to describe the new digital channels, such as specifying the program number as well as channel frequency. In addition, as electronic program guides were developed, additional information was sent to the STT to carry guide data. The format of the descriptions is unique to each vendor.

When terrestrial digital broadcast was introduced, the ATSC had to standardize a single descriptive format and include it in each system multiplex, because there is no equivalent of an out-of-band data channel for terrestrial broadcast. ATSC standard A/65 is called Program and System Information Protocol (PSIP).[41] The following tables are included in PSIP and shown in Figure 3.25.

System time table (STT) — provides the system time of day.

Master guide table (MGT) — defines other, optional tables, for example the EIT and ETT tables.

Virtual channel table (VCT) — provides information about the virtual channels (programs) contained in the system multiplex.

Rating Region Table (RRT) — provides a definition of the content advisory ratings for the system multiplex.*

All of these tables are transmitted on a well-known PID known as the *base PID*, which is defined as 0x1FFB for PSIP. Additional tables can be defined on separate PIDs, the event information table and extended text table are used to transmit further information about the programs in the system multiplex:

Event information table (EIT) — lists upcoming events that will be broadcast for each virtual channel (program) for the next 3 hours. Up to 128 EITs can be sent, each on its own PID, providing upcoming information for up to 16 days. The first four EITs are mandatory.

Extended text table (ETT) — lists descriptive information about each virtual channel (program). Just like the EITs, they have a 3-hour window, and up to 128 ETTs may be sent. However, they are optional.

Now that we have described the system information for broadcast, we can discuss the methods for cable systems. The ATSC assumed that cable systems, as rebroadcasters of off-air content, would naturally follow the A/65 standard. However, much of the programming on cable is no longer rebroadcast, and cable had already developed proprietary mechanisms for the carrying and coding of navigation information. Nevertheless, a compromise had to be worked out if it was ever going to be possible for a manufacturer to build a cable-ready digital television (CR-DTV). (See Section 3.7.4)

The approach was to take out-of-band information and make it available to the digital TV via the POD–host interface (see Chapter 21). The format is very

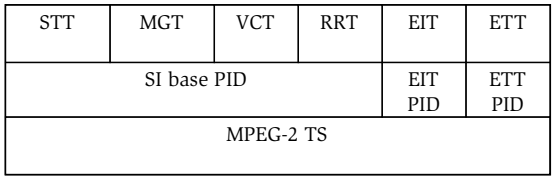


Figure 3.25 Protocol Stack for In-Band System Information. (From SCTE 40 2001, used with permission)

* The RRT for the United States is "hardwired" for two reasons: (1) It doesn't follow the linear "scale" of the original design and thus is a special case, and (2) since it is hardwired, the actual transmission of the RRT is not needed unless changed. Other regions can have downloaded RRTs and dynamic content advisory descriptors to match.

similar to the A/65 standard and allows an operational migration path to full A/65 compliance over time. (See ANSI/SCTE 65 Annex C for more details.)

Timing Distribution

The MPEG-2 transport stream carries timing information by means of a time stamp called the *program clock reference* (PCR). A PCR is placed in the adaptation header frequently enough (typically 10 times/second) to allow a phase-locked loop to recover the system time base at the receiver.

Video and audio PESs contain their own timing information, in this case presentation and decode time stamps. These are used by the decoder to synchronize the audio to the video.

Conditional Access

Although the MPEG-2 systems layer does not specify any conditional access (CA) mechanism, it does provide "hooks" for including conditional access control information in the system multiplex. Conditional access is not required and may be omitted if the multiplex is to be sent in the clear.

If the MPEG-2 system multiplex includes a conditional access table (CAT), it is transmitted continuously on PID value 1. The CAT contains an entry for each program within the multiplex, and the PID of the entitlement management message (EMM) stream for the CA system. (In cable systems, EMMs are typically sent out of band, and so this PID is usually null.)

In addition, the PMT may contain the PID for an entitlement control message (ECM) stream for the program. ECMs contain encrypted control words for the decryption circuitry (see Chapter 21).

Figure 3.26 illustrates how a receiver uses the CAT and PMT to obtain the information it needs to descramble the program.

Splicing

In a compressed video, bitstream cutting, or splicing, from one feed to another is not trivial. There are two main reasons for this.

1. The group of pictures (GOP) structure (see Figure 3.16) defines the frame order for a particular compressed bitstream. It defines which frames (or pictures) form a reference for predicted frames. If the bitstream is changed at an arbitrary point, these references will no longer make sense.
2. The encoder maintains a model of the decoder buffer fullness and codes each frame within a bit limit to ensure that the decoder buffer stays within prescribed bounds. If the bitstream is changed, the buffer will likely overflow or underflow.

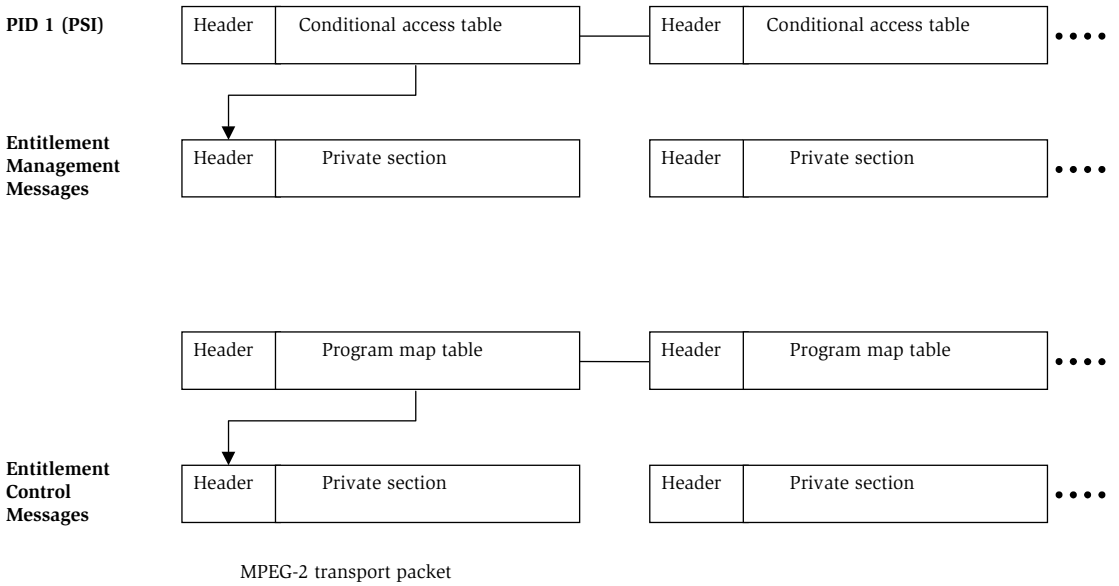


Figure 3.26 Conditional Access Table. (From [42] used with permission)

The Society of Motion Picture Experts (SMPTE) have defined mechanisms to establish a splice point in an MPEG-2 transport stream such that a seamless splice can occur.[43] At a splice point, the first frame of the new feed will be an I frame, and the GOP structure will specify that the predicted frames of the old feed should not use it as a reference frame. In addition, the buffer fullness is defined at the splice point so that the new feed can be encoded to fit within those constraints. For a predetermined splice, a fade to black at the splice point makes the transition much easier because there is no information to encode; still, it is possible to do a seamless splice of full-motion video. (However, note that this is not the method adopted by cable for digital program insertion — see Section 3.5.4).

Similar mechanisms exist for audio bitstreams, but splicing is much less complicated and audio muting is usually employed to cover any hiccups due to the transition.

3.5.2 Statistical Multiplexing

The goal of digital video compression is to reduce wasteful redundancy in the programming signal. Since programs have different amounts of redundancy at any given time and since the amount of redundancy varies over time in a given program, the compressed bitstream will have an instantaneously variable bit

rate. To accommodate this, the receiver needs to have sufficient buffer memory to smooth the flow of data to the decoders. This introduces a latency to accommodate buffer filling and emptying, and it requires standardization on the minimum buffer size. Although the buffer memory allows some flexibility, the problem of managing the flow of data is not completely eliminated. To ensure that the receiver buffer never underflows or overflows, the encoder maintains a model of the receiver buffer fullness and regulates the encoding rate so that the receiver buffer fullness always remains within design limits.

When the capacity of the channel compared with the average data rate of the programming allows multiple programs to be carried, an opportunity results. When there are multiple programs, the likelihood is that there will be a distribution of complexities among the channels that is amenable to statistical analysis. This is a reapplication of a phenomenon applied by long-distance phone companies decades ago. It was determined that much of conversation involves pauses when nothing is being said. If a large number of conversations are combined, the dead time of one conversation could be used to carry other conversations. If half a dozen conversations are combined, there is sufficient statistical dead time for the addition of one or two more conversations. Efficiency is dramatically increased. The scarce resource of transmission capacity is better utilized. The larger the number of combined conversations, the better the statistics and the better the results. This process is called *statistical multiplexing*. It applies equally well to video signals that have varying amounts of relatively dead time. The dead time in video is the time when little changes from frame to frame and the need for transmission capacity is minimal.

In practice, excellent results can be obtained with statistical multiplexing of video if there is a reasonably large number of channels within the system multiplex. For example, 10 constant-bit-rate 3.5-Mbps video bitstreams may be replaced by 14 video bitstreams with the same total instantaneous bit rate of 35 Mbps, producing a statistical gain of 40% with no perceived reduction in quality. However, it is important that the video bitstreams be uncorrelated; that is, they cannot all switch to challenging sports material at the same time.

Buffer memories for each of the multiplexed bitstreams facilitate the process of statistical multiplexing but require sophisticated algorithms to manage their utilization. As a buffer memory reaches its capacity, provisions must be made for it to be relieved. The better the algorithm, the more effective the result. The management of the buffer memories can be done at the origination point. As better algorithms become available, better results can be obtained using the existing receiving hardware.

Statistical multiplexing efficiency can be optimized if a bank of real-time video encoders is combined with a statistical multiplexer. In this arrangement, each encoder dynamically requests a bit rate according to the complexity of material at that point and a specified quality level (as set by an operator). The statistical multiplexer will dynamically assign a target bit rate to each encoder, depending on (a) the bit rate requested by it and (b) the sum of the bit rates

requested by all encoders. In this way, each video encoder will achieve the best quality it can, given the constraints established by the multiplexer. In addition, statistical multiplexing systems may incorporate "look-ahead" preprocessing to anticipate bit needs before they happen. Using these techniques, the multiplexer is able to use the entire capacity of the channel with no possibility of overflow.

3.5.3 Reencoding and Transcoding

Reencoding provides a way to dynamically change the encoding parameters of a video stream after it has been encoded. Various techniques have been developed, but fundamentally they all work by requantizing the DCT coefficients. In some cases a significant reduction in bit rate can be achieved with little or no loss of quality.

Transcoding can be used to convert a bitstream encoded in one format to another. This can be done by decoding to baseband video and then encoding according to the new format; however, this is expensive and causes delay. Worse still, the second encoder may struggle to encode artifacts created by the first because they contain significant random information (for example, blocking or mosquito noise). A well-designed transcoder can minimize these effects, and reduce delay, by recoding the bitstream without the need to decode all the way back to a baseband signal. Better results are possible because it knows about the source and destination formats. Transcoders are able to change resolution and even render a graphical overlay on the video.

Clearly, reencoding or transcoding is an important tool at any point in which compressed bitstreams are combined to create a system multiplex. Statistical multiplexing can be much more efficient because peaks that occur simultaneously across all bitstreams can be smoothed by judicious buffer management, requantization, or even transcoding. In addition, reencoders and transcoders maintain a model of the decoder buffer fullness and work within its constraints so that the receiver buffer will not underflow or overflow.

3.5.4 Digital Program Insertion

Local commercial insertion is an important source of revenue for the cable operator. It is simple to cut from one audio/video feed to another, and the practice of *local ad insertion* has been commonplace for many years, originally using videotape machines to play the local commercials and more recently using *digital ad insertion* servers. The time periods allotted for the insertion of commercials, known as *avails*, are signaled by cue tones, which are sent about 5–8 seconds before the commercial is to be inserted (allowing for *preroll* of the tape machine or server). (See also Section 8.9)

In a digital headend, equivalent functionality must be preserved; that is, it must be possible to insert a digitally compressed commercial into a MPEG-2 transport

stream. It is theoretically possible to decode all the programs in an MPTS to baseband video and audio, insert the commercial using traditional techniques, and then reencode and reassemble the MPTS. However, this approach is not recommended because it would be extremely expensive and cause a significant time delay and, very likely, a loss of quality. Therefore, an SCTE Digital Video Standards subcommittee has worked very hard for a number of years to develop standards for *digital program insertion* (DPI) based on splicing of compressed digital content[44,45].

The DPI standards do not specify exactly how the splicing of the compressed bitstream should be achieved (see Section 3.5.1), because there are a number of acceptable techniques, and the intention is to allow innovation on the part of the vendors. Encoding and/or transcoding methods are extremely useful to allow a smooth transition between one bitstream and another; in practice, many DPI splicers use these techniques to allow near-seamless splicing at an arbitrary point in the bitstream. Instead of specifying the splicer implementation, the DPI standards specify a standard *applications programming interface* (API) between the network and the DPI system. The reference model for DPI is shown in Figure 3.27.

The input to the DPI system, called the *network stream*, is an MPEG-2-compliant transport stream. Any programs that are to be spliced will include a *registration descriptor* in their PMT. In addition, a separate PID stream carries a *splice information table*, which contains the splice commands and the presentation time of the splice points for that program.

The splice subsystem intercepts and, if necessary, decrypts the splice commands before passing them to the local ad server. At an avail, the ad server plays out the MPEG-2 audio/video elementary streams and the splicer inserts them

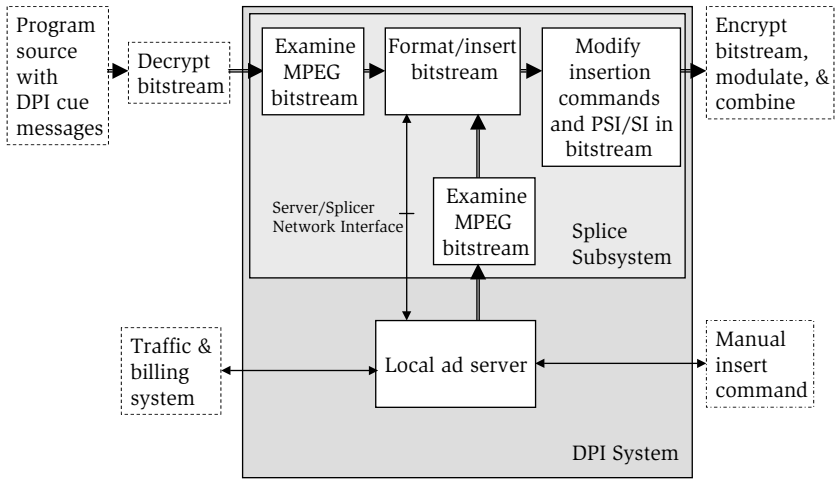


Figure 3.27 DPI Headend System Overview. (From [46] figure 1, page 5, used with permission)

into the network stream from the specified *splice in-point* until the specified *splice out-point*. An NCTA conference paper provides a good background on some of the implementation details of DPI for further reading.[47]

3.5.5 Emergency Alert System

Local emergency alerts are currently provided by a crawling text overlay along the bottom of the screen. Sometimes they are accompanied by audio tones. In other applications, text and graphics are overlaid on the video. During a national alert or certain significant local/regional events, a video and audio override of every channel is required. The common analog practice of source switching is relatively simple to implement, but it is much more expensive and complicated to do this with a digital multiplex.

The SCTE have defined an emergency alert message for cable to allow the alert event to be signaled to the receiver.[48] All digital receivers have the capability to render text messages over the video and to generate audio tones. In the unlikely event of a national emergency, the receiver can be instructed to force-tune to a channel for the duration of the alert. The emergency alert message is formatted as an MPEG-2 table that contains the following fields:

- Start time and expected duration of the emergency alert
- Text message description of the alert
- Availability and location of a details channel
- Force-tune flag
- A pointer to an audio channel to replace the audio for the duration of the emergency

The emergency alert message may be sent in-band within the system multiplex for in the clear channels, allowing it to be inserted by a terrestrial digital broadcaster for retransmission over a cable system. Alternatively, for scrambled channels, the emergency alert message can be transmitted in the out-of-band channel. In this case, it is received by the point of deployment (POD) module and relayed to the host as an MPEG-2 table (using the same mechanism previously described for navigation messages).

3.6 Digital Transmission

Now that we have described how an MPEG-2 system multiplex is created, we need to explore how it is delivered over an analog cable system. Moreover, the MPEG-2 transport stream assumes an error-free transmission path, and so forward error correction information must be added to the multiplex before it is modulated onto a carrier signal.

3.6.1 Modulation

Modulation serves the same purpose for digital signals as it does for analog signals. Although digital signals can conveniently utilize time division multiplexing, frequency division multiplexing (FDM) remains an important part of the plan. There are several reasons for this. Practically, digital circuits are limited in their speed. They cannot support a TDM structure for the entire gigahertz spectrum of cable. Breaking the spectrum into 6-MHz pieces relaxes the speed requirement on the digital circuits. TDM can then be utilized within each of the 6-MHz channels.

Additionally, FDM facilitates the transition from analog television to digital television. Cable systems have added digital for those customers who want more programming choices. They have added digital to the current collection of analog channels. In some cases, the analog scrambling is left in place, but it is also common to simulcast premium channels in the digital tier to avoid the cost of incorporating analog descrambling circuitry into the digital STT. By separating the spectrum into two areas, one for analog channels and one for digital channels (including TDM), a rational transition to digital technology has been accomplished.

As the cost of a digital STT continues to fall, it is becoming feasible to use digital compression as an alternative to a plant upgrade. For example, in a 450-MHz system, only must-carry channels are sent in analog format (maybe channels 2–40), and all other programming is sent in digitally compressed format. Each subscriber would need a digital STT, but the cost may be less than that for upgrading the system to 750 MHz.

3.6.2 Forward Error Correction (FEC)

One of the most important features of digital signals is the ability to judiciously add redundancies that allow error detection and correction. The understanding of error detection and correction begins with the realization that an error is the conversion of a 1 to a 0 or the conversion of a 0 to a 1. The fact that the error is limited to just those two cases is very powerful. Earlier in this chapter, we discussed the simplest of error-detection schemes, parity checking. If parity is taken into two dimensions, substantially more powerful results can be obtained.

Figure 3.28(a) shows seven binary words of seven bits each. In Figure 3.28(b), spaces are introduced to help identify the clusters of seven bits. In Figure 3.28(c), the seven binary words of seven bits each are arranged in a two-dimensional array. An even parity bit is attached to each row. Recall that even parity means that the appended bit will be a 1 if an odd number of 1s appear in the word and a 0 if an even number of 1s appear. Parity allows us to determine if an odd number of errors has occurred in any of the words, but it does not allow us to determine where the error occurred.

Now let us form parity bits in the vertical direction as well. This will give us an eight-bit byte at the bottom of the array. We now have 8 columns and 8 rows, for 64 bits. This array can be strung out into a serial stream, as shown in Figure 3.28(e).

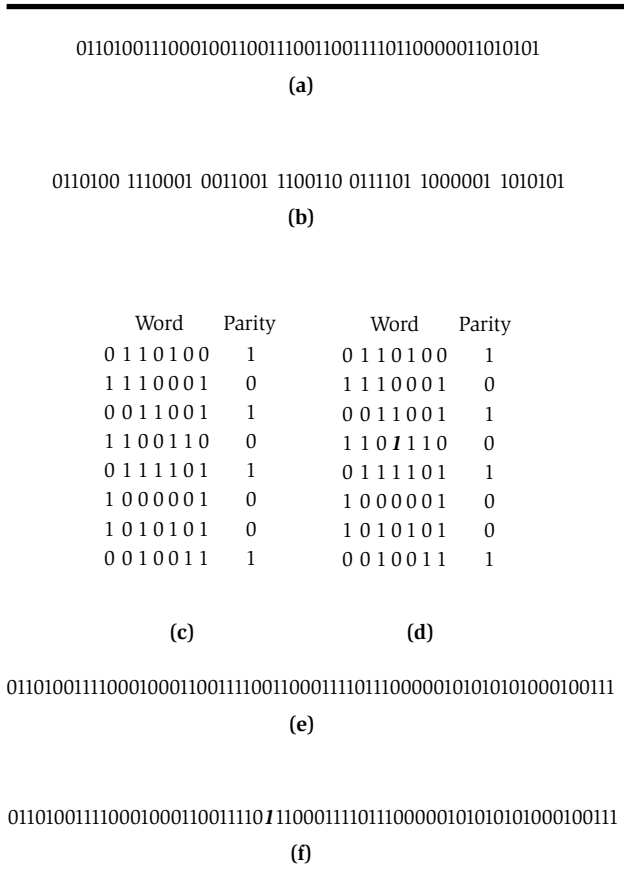


Figure 3.28 Block code example.

If an error occurs, such as shown in Figure 3.28(f), it will convert a 0 to a 1 or it will convert a 1 to a 0. In this example, a 0 has been converted into a 1. If at the receive site, we rearrange the received bits into a block as shown in Figure 3.28(d) and apply the parity-checking principle to the rows and columns, we will find a violation in row 4 and column 4. This tells us we have an error and it also tells us where! Since an error can only be an inversion of the logic, we can correct the error perfectly. A little experimentation with this example will reveal that if we have two or more errors that do not occupy the same row or column, we can locate them and correct them. This would work for up to seven errors as long as they were not in the same row or column. If, however, two errors occurred in the same row (or the same column), we could determine that we have two errors, but we could not locate the row (column) in which they occurred. Knowing that two errors exist is valuable information.

In this example, a penalty is paid in complexity. Thirty percent more information is sent. In addition, the hardware at the sending and receiving end is more complex. This example is termed a *block code*, since the data is partitioned into blocks for processing. More complex codes exist that are beyond the scope of this text to explain. They involve advanced mathematical concepts and are very effective.

Several error-combating strategies are often combined in digital transmission systems. A most important error-combating strategy is called *interleaving*. Figure 3.29 shows how this works. The goal is to take adjacent bits and spread them out in time so that a noise burst does not destroy adjacent bits. When the

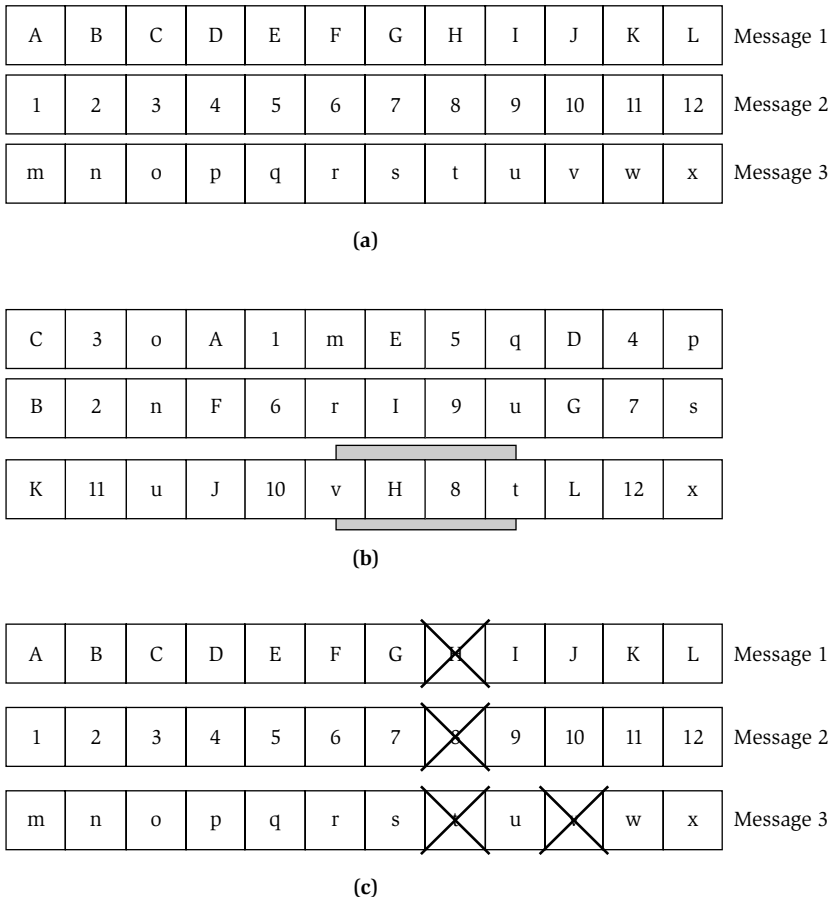

A B C D E F G H I J K L Message 1
1 2 3 4 5 6 7 8 9 10 11 12 Message 2
m n o p q r s t u v w x Message 3
(a)
C 3 o A 1 m E 5 q D 4 p
B 2 n F 6 r I 9 u G 7 s
K 11 u J 10 v H 8 t L 12 x
(b)
A B C D E F G I J K L Message 1
1 2 3 4 5 6 7 9 10 11 12 Message 2
m n o p q r s u w x Message 3
(c)

Figure 3.29 Interleaving to protect against burst errors. (a) Order of interleaving. (b) Interleaved messages suffering a burst of noise. (c) De-interleaved messages suffering nonadjacent losses.

interleaving is undone in the receiver, the group of destroyed bits is distributed over a length of time, so damaged bits are not adjacent. Then techniques such as the block code example can detect and correct the errors. Interleaving converts bursty noise into something more like random noise.

3.6.3 Required Signal Quality

Digital systems have a "cliff effect" that allows them to have perfect or near-perfect pictures as the received signal weakens, until a point is reached where the picture fails altogether. It happens very abruptly. All viewers have essentially the same quality picture and sound, no matter where they are in the service area. When the signal strength falls below a threshold, the picture essentially disappears.

Noise in an analog picture appears as snow and begins to mask the image and get in the way of the enjoyment of the picture. If the programming is compelling, viewers will put up with this and continue watching even very poor pictures. If the programming is marginal, viewers will switch to something else. Viewers of digital images won't get the chance to make that decision. The picture will be just as noise free at the home as in the studio. When the noise becomes excessive, the picture will be useless. It will fall off the cliff.

Multipath distortions (ghosts) that degrade an analog picture either destroy the digital picture by pushing it over the threshold or have no impact at all. The digital broadcast system's built-in multipath canceller system brings reception to areas not previously served.

Field tests in North Carolina have indicated that the broadcast digital system is so robust that anywhere an analog picture was even marginally acceptable, the digital picture was perfect. In many cases where the analog picture was unwatchable, the digital picture came in just fine.

Cable spectrum is much better behaved than the over-the-air broadcast media. Either the ghost canceller is not needed or its design parameters are much more relaxed as compared with the broadcast case. If the analog signal is minimally acceptable, the digital signal will provide essentially perfect pictures (to the extent that the compression quality is acceptable).

3.6.4 Picture Quality

Picture quality is a function of the residual digital artifacts. These are determined primarily by the aggressiveness of the compression, the detail and motion in the picture, and the quality of the encoder.

Although expert viewers can easily spot digital artifacts in currently available digital video, there is little to no complaint from consumers. Standard-definition movies encoded at 3 Mbps using MPEG-2 MP@ML provide excellent results (by comparison, the average bit rate of a DVD is 4 Mbps). Movies can be encoded in high definition at 12 Mbps at MPEG-2 MP@HL with almost no loss of quality.

Video can be more challenging, especially high-motion material such as sports events. Variable-bit-rate compression combined with statistical multiplexing is commonly used to keep up with rapid motion while maintaining an efficient multiplex. To compress difficult sports material up to 6 Mbps may be required during peaks, even with a good MPEG-2 encoder. However, an average rate of 3 Mbps will probably suffice for a system multiplex containing a range of video material (sports, news, current affairs, and so on).

3.7 Digital Television Standards

Standards are of vital importance to the deployment of digital television. A digital STT is extremely complex and can be manufactured cost-effectively only in very high volumes. Moreover, standards support the development of high-volume chipsets that can be sold to a worldwide marketplace.

In a terrestrial broadcast environment, a standard format for digital television is required that is understood by all receivers that can receive the signal. Hence, the ATSC standard had to be developed in advance of deployment. In cable (and satellite) environments, proprietary systems were developed initially, but even in this case common standards (MPEG-2 video compression and transport) have been used to build cost-effective systems.

3.7.1 ATSC

The Advanced Television Systems Committee (ATSC) has developed a Digital Television Standard[49] that is a complete system for terrestrial broadcast of high-quality video, audio, and ancillary data. It defines the following.

1. Source coding and compression: MPEG-2 video compression (standard and high definition) and AC-3 audio compression (A/52).
2. Service multiplex and transport: The MPEG-2 systems layer is used for multiplexing and transport.
3. RF/transmission is 8-VSB (terrestrial mode) or 16-VSB (high-data-rate mode).
4. Navigation of program content within the service multiplex is defined by the Program and System Information Protocol (A/65).

For further information, the reader is referred to the *Guide to the Use of the Digital Television Standard*.[50]

3.7.2 SCTE

The Society of Cable Telecommunications Engineers (SCTE) is an accredited American National Standards Institute (ANSI) organization that utilizes the due

process methods required by ANSI to set standards for the cable industry. This is the mechanism by which digital video for use on cable is standardized. These SCTE standards are an integral part of OpenCable.

The *Digital Video Systems Characteristics Standard* covers the subset of MPEG-2 video compression adopted by cable.[51] The *Digital Video Transmission Standard* for cable television describes the framing structure, channel coding, and channel modulation for a cable television realization of a digital multiservice distribution system.[52] The standard applies to signals that can be distributed via satellite directly to a cable system, as is the usual cable practice. The specification covers both 64-QAM and 256-QAM.

The modulation, interleaving, and coding are tailored to the characteristics of cable system practice in North America. The modulation is selected by the cable system to be either 64-QAM or 256-QAM. The forward error-correction approach utilizes a concatenated coding approach that produces low error rates at moderate complexity and overhead. The system goal is one error event per 15 minutes. The input data format is MPEG-2 transport packets, which are transmitted serially with the MSB first.

Two modes are supported: Mode 1 for use with 64-QAM with a symbol rate of 5.057 Msps, and Mode 2 for use with 256-QAM with a symbol rate of 5.361 Msps.

The cable channel, including optical fiber, is a bandwidth-limited linear channel, with a balanced combination of white noise, interference, and multipath distortion. The QAM technique used, together with adaptive equalization and concatenated coding, is well suited to this application and channel.

3.7.3 OpenCable

CableLabs, the cable industry's research and development consortium, has created a number of purchasing specifications on behalf of the industry. These are not standards, since CableLabs is not an accredited standards-setting organization. Rather, these specifications define the functionality and external parameters of products used in cable but leave the internal details to the manufacturers. For example, CableLabs will not specify a single microprocessor or operating system for these applications. Nevertheless, to ensure interoperability, CableLabs establishes a certification program to verify that the products comply with the specification and will perform satisfactorily.

CableLabs announced a "Harmony Agreement" for digital standards in October 1997. Five open-layer standards are included in the agreement: MPEG-2 video compression, Dolby Digital (AC-3) audio system, MPEG-2 transport, ATSC system information, and the International Telecommunications Union ITU-J.83 Annex B modulation system. The modulation system calls for 64-QAM carrying 27 Mb/s and 256-QAM carrying 38.8 Mb/s with concatenated Trellis code plus enhancements such as variable interleaving depth for low latency in delay-sensitive applications. All these agreements have been endorsed by CableLabs

and have since been adopted as digital standards by the Society of Cable Telecommunications Engineers.

The goal for OpenCable is to create a multivendor interoperable environment of digital set-top terminals. An important objective is the competitive supply of all the elements of the system, including the advanced set-top terminal, its operating system, core applications and plug-ins, conditional access system, platform management system, and billing, provisioning, and subscriber management systems.

The procedure began with a request for information (RFI) sent to leading computer and consumer electronics companies. Twenty-three companies responded. From these submissions, CableLabs solicited vendors to author specific elements of the suite of specifications. An intellectual property (IP) pool was created with cross-licensing of all IP required for the specification. Implementation IP not inherently required by the specification was left to vendors to license in their normal fashion.

Key OpenCable interfaces include the formats for digital cable television, consumer privacy systems, copyright protection system, interfaces for high-speed connection to the Internet, and interfaces required to author interactive applications. The last of these is the OpenCable Applications Platform (OCAP), which is designed to allow the cable operator to download applications to retail STTs in order to provide a seamless environment for the cable subscriber. OCAP is a middleware specification that enables application portability across a wide range of home devices and cable networks. OCAP 1.0 defines a Java-based execution engine (EE), and OCAP 2.0 extends that platform with the addition of Web-based technologies such as XHTML, XML, and ECMAScript.

CableLabs will implement a vendor's certification process, whereby specific products can be tested for compliance with the OpenCable specifications. The certification process will allow these devices to be made available to consumers through retail outlets. CableLabs recommends that its members utilize the specifications for their procurements. However, members will make their own procurement decisions. This is not a joint purchasing agreement. The OpenCable specifications are available through the OpenCable Web site (www.opencable.com).

3.7.4 Digital Cable Compatibility

There has been a lot of interest in making "digital cable-compatible" devices available at retail. Cable operators and consumer electronics manufacturers both stand to gain, and the FCC believes that competition among manufacturers will provide more and better choices to the consumer.

A memorandum of understanding (MOU) between eight larger MSOs and 14 consumer electronics manufacturers was sent to the FCC in a letter to the chairman on Dec. 19, 2002. Although the memorandum covers only "unidirectional digital cable products," compliance with a significant number of standards is required to successfully tune, demodulate, descramble, decode, and display a

digital program. As already discussed, digital services tend to be "all or nothing"; that is, they work perfectly or not at all!

Although the MOU is just a first step toward cable-ready digital televisions, it is worth summarizing the list of SCTE standards that are required to support plug-and-play compatibility.

- The Cable Network Interface specification[53] establishes the interface between the cable plant and the cable-compatible receiver. Building on the Harmony Agreement, it includes RF transmission, modulation, video and audio compression, navigation, conditional access, Emergency Alert System (EAS) messages, and the carrying of closed captioning. Many of these specifications have already been discussed in this chapter and will be examined further in Chapter 21.
- The Digital Video Service Multiplex and Transport System for Digital Cable specification[54] defines all of the various operating parameters related to the MPEG-2 transport stream, including various constraints designed to ensure interoperability.
- The Service Information Delivered Out-of-Band for Digital Cable Television specification[55] describes how navigation information is sent over the cable system so that a receiver can display a list of available programs to the user and allow program selection.
- The Host–POD Interface specification[56] defines a standard interface between the cable-compatible receiver and a replaceable conditional access module (called a point-of-deployment, or POD, module). The cable operator will supply POD modules to subscribers to allow them to view premium digital channels. (See Chapter 22 for more details.)
- The POD Copy Protection System specification[57] defines a method of copy protection so that premium digital content passed across the POD–Host interface is protected from piracy. (See Chapter 22 for more details.)

The principle behind certification of cable-ready devices is primarily that they should do no harm to the network. A prototype receiver will be examined by CableLabs, but thereafter the consumer electronics manufacturers will self-certify their products. The provisions of the MOU are scheduled to take effect in 2004 for 750-MHz (or greater) cable systems, and further work is still required to enable advanced services such as enhanced program guides, video on demand, and interactive television applications.

3.8 Summary

We have discussed the fundamentals of digital compression technology for video and audio as they apply to cable television. We presented the discrete cosine

transform and its inverse transform. The DCT is applied to still images, and differential techniques are utilized to reduce redundancy between frames. For audio signals, sub-band and/or frequency transform coding may be used, combined with psychoacoustic masking, to substantially reduce bit rate while maintaining good perceptual quality.

Audio and video bitstreams are multiplexed into a single transport stream for distribution over the cable system. The transport stream includes information that describes the system multiplex, allowing the receiver to select and decode a single program. Error-detection and -correction techniques are used to maintain the integrity of the transport stream in the presence of a noisy transmission environment.

The application of digital techniques to cable dramatically increases the number of programs and the quality of the delivered signals while facilitating new services.

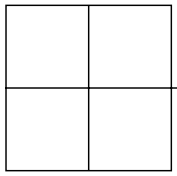
Endnotes

1. Chen and Pratt, Scene Adaptive Coder, *IEEE Transactions on Communications*, March 1984.
2. Joel Brinkley, *Defining Vision: The Battle for the Future of Television*. New York: Harcourt Brace, 1997, pp. 122–128.
3. Ibid., p. 116.
4. MPEG, *Information Technology — Generic Audio Coding of Moving Pictures and Associated Audio, Part 2: Video*. International Standard IS 13818-2, ISO/IEC JTC1/SC29 WG11, 1994.
5. Jerry D. Gibson et al., *Digital Compression for Multimedia — Principle and Standards*. San Francisco: Morgan Kaufmann, 1998.
6. Justin Junkus, *DigiPoints*, Vol. 2.: Society of Cable Telecommunications Engineers, 2000, Chap. 8.
7. Jerry Whitaker, *DTV: The Revolution in Electronic Imaging*. New York: McGraw-Hill, 1998, Chap. 5.
8. H. Nyquist, Certain Factors Affecting Telegraph Speed. *Bell System Technical Journal*, 3(1924):324–346.
9. H. Nyquist, Certain Topics in Telegraph Transmission. *AIEE Transactions*, 47(1928):617–644.
10. Jerry Whitaker, *DTV: The Revolution in Electronic Imaging*. New York: McGraw-Hill, 1998, p. 123, 189.
11. ITU-R BT.601-5, *Encoding parameters of digital television for studios*.
12. Mischa Schwartz, *Information Transmission, Modulation, and Noise*, 2nd ed. New York: McGraw-Hill, 1970.

13. S. Merrill Weiss, *Issues in Advanced Television Technology*. Boston: Focal Press, 1996.
14. IEEE, *Standard Specification for the Implementation of* 8 × 8 *Inverse Discrete Cosine Transform*, std. 1180–1990, Dec. 6, 1990.
15. Charles Poynton, *A Technical Introduction to Digital Video*. New York: Wiley, 1996.
16. K. Sayood, *Introduction to Data Compression*. San Francisco: Morgan Kaufmann, 1996.
17. D. A. Huffman, A Method for the Construction of Minimum Redundancy Codes, *Proceedings of the IRE*, Vol. 40, Sept. 1952, pp. 1098–1101. Institute of Radio Engineers, New York.
18. A. N. Netravali and B. G. Haskell, *Digital Pictures, Representation and Compression*. New York: Plenum, 1988.
19. EIA-608-B (2000), *Line 21 Data Services*.
20. ANSI/SCTE 21 2001 (formerly DVS 053), *Standard for Carriage of NTSC VBI Data in Cable Digital Transport Streams*, available at www.scte.org.
21. ANSI/SCTE 20 2001 (formerly DVS 157), *Standard Methods for Carriage of Closed Captions and Non-Real-Time Sampled Video*, available at www.scte.org.
22. EIA-708-B (1999), *Digital Television (DTV) Closed Captioning*.
23. SMPTE 240M (1995), *Standard for Television — Signal Parameters — 1125-Line High-Definition Production Systems*.
24. SMPTE 296M (1997), *Standard for Television*, 1280 × 720 *Scanning, Analog and Digital Representation, and Analog Interface*.
25. SMPTE 274M (1995), *Standard for Television*, 1920 × 1080 *Scanning and Analog and Parallel Digital Interfaces for Multiple-Picture Rates*.
26. ANSI/SCTE 43 2001 (formerly DVS 258), *Digital Video Systems Characteristics Standard for Cable Television*, available at www.scte.org.
27. MPEG, *Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to 1.5 Mbit/s, part 2: Video*. International Standard IS 11172-2, ISO/IEC JTC1/SC29 WG11, 1992.
28. Michael Adams, *OpenCable Architecture*. Indianapolis, IN: Cisco Press, 2000, Chap. 9.
29. J. Watkinson, *Compression in Video & Audio*. Boston: Focal Press, 1995.
30. K. Brandenburg and M. Bosi, *Overview of MPEG-Audio: Current and Future Standards for Low-Bit-Rate Audio Coding*. 99th Audio Engineering Society Convention, New York, Preprint 4130, 1996.
31. Jerry D. Gibson et al., op. cit., Chapter 8.
32. Jerry Whitaker, op. cit.

33. ATSC, *Digital Audio Compression Standard (AC3)*. Advanced Television Systems Committee, Washington, DC, Doc. A/52, Dec. 20, 1995.
34. MPEG, *Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to 1.5 Mbit/s, Part 3: Audio*. International Standard IS 11172-3, ISO/IEC JTC1/SC29 WG11, 1992.
35. MPEG, *Information Technology — Generic Audio Coding of Moving Pictures and Associated Audio, Part 3: Audio*. International Standard IS 13818-3, ISO/IEC JTC1/SC29 WG11, 1994.
36. MPEG, *MPEG-2 Advanced Audio Coding, AAC*. International Standard IS 13818-7, ISO/IEC JTC1/SC29 WG11, 1997.
37. MPEG, *Information Technology — Generic Audio Coding of Moving Pictures and Associated Audio, Part 1: Systems*. International Standard IS 13818-1, ISO/IEC JTC1/SC29 WG11, 1994.
38. Jim Taylor, *DVD Demystified*. New York: McGraw-Hill, 1998, p. 135.
39. Michael Adams, *OpenCable Architecture*. Indianapolis: Cisco Press, 2000, p. 64.
40. Michael Adams, *OpenCable Architecture*. Indianapolis: Cisco Press, 2000, p. 65.
41. ATSC, *Program and System Information Protocol for Terrestrial Broadcast and Cable*. Advanced Television Systems Committee, Washington, DC, Doc. A/65, Dec. 23, 1997.
42. Michael Adams, *OpenCable Architecture*. Indianapolis: Cisco Press, 2000, p. 66.
43. SMPTE 312M (1999), *Standard for Television, Splice Points for MPEG-2 Transport Streams*.
44. ANSI/SCTE 30 2001 (formerly DVS 380), *Digital Program Insertion Splicing API*, available at www.scte.org.
45. ANSI/SCTE 35 2001 (formerly DVS 253), *Digital Program Insertion Cueing Message for Cable*, available at www.scte.org.
46. ANSI/SCTE 67 2002 (formerly DVS 379), *Applications Guidelines for SCTE 35 2001*, available at www.scte.org.
47. M. Kar, S. Narasimhan, and R. Prodan, Local Commercial Insertion in the Digital Headend, *Proceedings of NCTA 2000 Conference*, New Orleans.
48. SCTE 18 2002 (formerly DVS 208), *Emergency Alert Message for Cable*, approved as a joint standard with CEA as ANSI-J-STD-042-2002, available at www.scte.org.
49. ATSC, *Digital Television Standard*, Advanced Television Systems Committee, Washington, DC, Doc. A/53, Sept. 16, 1995.
50. ATSC, *Guide to the Use of the Digital Television Standard*, Advanced Television Systems Committee, Washington, DC, Doc. A/54, Oct. 4, 1995.

51. ANSI/SCTE 43 2001 (formerly DVS 258), *Digital Video Systems Characteristics Standard for Cable Television*, available at www.scte.org.
52. ANSI/SCTE 07 2000 (formerly DVS 031), *Digital Video Transmission Standard for Television*, available at www.scte.org.
53. SCTE 40 2001 (formerly DVS 313), *Cable Network Interface Specification*, available at www.scte.org.
54. ANSI/SCTE 54 2002 (formerly DVS 241), *Digital Video Service Multiplex and Transport System for Cable Television*, available at www.scte.org.
55. ANSI/SCTE 65 2002 (formerly DVS 234), *Service Information Delivered Out-of-Band for Digital Cable Television*, available at www.scte.org.
56. ANSI/SCTE 28 2001 (formerly DVS 295), *Host–POD Interface*, available at www.scte.org.
57. ANSI/SCTE 41 2001 (formerly DVS 301), *POD Copy Protection System*, available at www.scte.org.



Chapter 4
Digital Modulation

4.1 Introduction

This chapter does not claim to be a treatise on data communications. Rather, it offers concepts that are particularly germane to the practice of data transmission on HFC networks. It is intended primarily to be an introduction to data communications for cable practitioners who have not previously been exposed to the subject. The examples of modulation methods and communications protocols are restricted to those that are most relevant to cable data communications.

Section 4.2 discusses elementary concepts in digital modulation. It is intended for engineers who may be familiar with analog modulation as practiced in the cable television industry but who have not previously worked with digital modulation. The section discusses the digital modulation formats most germane to cable television.

Section 4.3 introduces the concepts of multiplexing, a way of sharing the spectrum among many users. Section 4.4 introduces methods of measuring the quality of digitally modulated signals.

4.2 Modulation Technology

The object of data transmission is to get a stream of bits from here to there. *Bit*, which stands for *bi*nary digi*t*, is a voltage that can take on only one of two levels. We arbitrarily call these levels 1 and 0. Often, but not always, they are represented in electronic circuits by the voltages +3.3 (or +5) and 0 volts. The higher voltage usually, but not necessarily, represents a logical 1, while 0 volts represents a logical 0. A combination of many of these bits make up *digital data*.

It is possible to transmit bits of data over an RF carrier using the same variables as are used in analog transmission. One can vary, or "modulate," the amplitude, frequency, or phase of the carrier with the information to be transmitted. Frequency modulation is often used for very simple data communications links, due to the ability to construct low-cost FM receivers capable of very good performance. However, frequency modulation does not provide good "spectral efficiency."

4.2.1 Spectral Efficiency

One figure of merit used to compare different modulation formats is *spectral efficiency*, expressed in bits per hertz: the number of bits per second of data, divided by the RF bandwidth required to transmit it. A modulation type with high spectral efficiency can transmit more bits of data per second in the same number of hertz of RF bandwidth. On the other hand, it can be expected to cost more and be less *robust* (more susceptible to noise and transmission distortion). For example, we shall show that the theoretical spectral efficiency of biphase shift keying (BPSK) modulation is 1 bit per hertz. That is, when one is using BPSK modulation, the theoretical bandwidth required to transmit the signal is equal to the bit rate: a 1-Mb/s (one megabit per second) data signal will require 1 MHz of RF bandwidth. On the other hand, if the same signal were transmitted using 256-QAM, the theoretical bandwidth required is only 0.125 MHz because the spectral efficiency is 8 bits per hertz ($1/8 = 0.125$). However, all else being equal, 256-QAM requires a carrier-to-noise ratio (C/N) more than 13 dB better, and the cost of the hardware is considerably higher. In all cases, the actual occupied bandwidth is wider than the theoretical because of practical considerations.

4.2.2 Frequency Shift Keying (FSK)

The simplest commonly used modulation method is frequency modulation, called frequency shift keying, or FSK. Information is transmitted by shifting an oscillator between two frequencies. The receiver consists of a conventional frequency demodulator. It is followed by a shaping circuit, which restores the pulses to their original shape.

FSK systems tend to be popular for relatively low-speed data transmission where receiver cost is paramount. Applications include early-generation downstream data transmission to addressable set-top converters, upstream transmission from some early RF impulse pay-per-view (IPPV) converters, and status monitoring. Spectral efficiency is very poor, but this is relatively unimportant in low-speed applications. The transmission is robust (not easily damaged by noise or distortion), and the required hardware is quite low in cost. Single-chip data receivers have been available for many years.

Figure 4.1 illustrates a simple FSK communications system. Data, in the form of a series of 1s and 0s, is applied to a low-pass filter whose function is to minimize the number of sidebands produced by the modulation process. The filtered waveform is applied to a frequency modulated oscillator, whose frequency deviation is proportional to the voltage applied at the modulating input. *Frequency deviation* is the difference between an instantaneous frequency and the rest, or carrier, frequency. A parameter of an FSK system is the peak-to-peak deviation, or the difference between the two frequencies

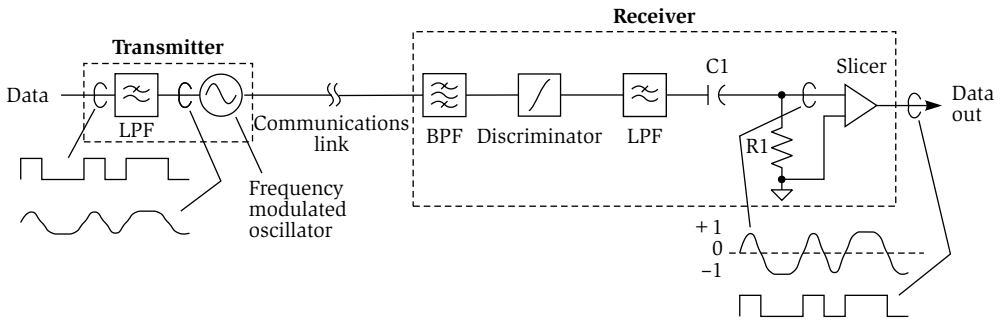


Figure 4.1 A simple FSK system.

of the oscillator when 1 and 0 are applied.* The wider the deviation, the more immune the system is to noise but the more spectrum the signal occupies.

After transmission to the receiving end, the signal is usually passed through a bandpass filter, BPF, and then applied to a discriminator, which is very similar to one used for analog signal recovery. A low-pass filter, LPF, is usually used after the discriminator, to reduce out-of-band noise picked up during transmission. Capacitive coupling is shown after the filter, to make sure the signal can be "sliced" at its average value. From the capacitor, the signal is supplied to a data slicer, a comparator whose output is 1 when the input voltage is above ground and 0 when it is below ground (or vice versa). The slicing process restores the wave shape of the original signal.

Most transmission systems require that the dc value of the data waveform be removed for transmission. This means that, in some manner, the waveform averaged over some reasonable interval must have a value of 0 volts. That condition is violated if the datastream contains an excessively long string of 1s or 0s. If the signal doesn't have the same quantity of 1 and 0 states when averaged over some suitable time interval, C1 will take on a charge and the slicing will not occur in the center of the waveform. Later we show several ways to ensure that a signal doesn't have a dc component.

---

* In the analog world, deviation is normally measured on a peak basis — the difference between either extreme frequency and the center frequency. Peak-to-peak deviation is twice the peak deviation. In digital transmission, the carrier never rests at the center frequency, so peak deviation tends to have less meaning. One must be precise, however, in distinguishing between peak and peak-to-peak deviation.

Filtering in FSK Transmission Systems

At least three stages of filtering are used in the transmission system. A low-pass filter at the transmitter limits the spectrum of the baseband signal (before modulation). In many other modulation methods, the filter may be implemented either at baseband or at RF, after modulation. However, in FSK the nonlinear nature of frequency modulation precludes doing so. The receiver contains a bandpass filter at RF (or intermediate frequency, IF), used to limit the energy applied to the discriminator. After demodulation, a baseband low-pass filter is used to remove as much noise as possible.

In low-cost FSK systems, the filtering philosophy often recognizes that either the transmitter or the receiver is in the subscriber's home, and this end must be built at minimum cost. Placing most of the filtering at the headend helps keep the cost of the home equipment down. This somewhat compromises the noise performance or bandwidth efficiency in favor of low cost. For example, a data communications system that communicates from the headend to addressable set-top converters will employ significant filtering at the transmitter. At the receiver, the IF filter frequently employs a low-cost ceramic filter made for FM receiver use (180- to 200-kHz bandwidth), and the deviation is set to make the occupied spectrum match the width of the filter.

Manchester Encoding

We now describe one method for adding a clock to the data signal and for removing any dc component. The method is independent of the modulation but is introduced at this point because it is often used with FSK modulation. Manchester encoding is also used in 10/100BaseT Ethernet. In data transmission, it is necessary not only to recover the transmitted data but also to recover a clock signal, used to define when a data state should be sampled. A clock is necessary in order to allow the receiver to "know" when a new bit arrives. The clock must be recovered from the transmitted data, which can be done in many ways. Manchester encoding is often used in low-cost systems due to its simplicity. The data is said to be "self-clocking." The penalty is that the spectral efficiency of the transmission is halved compared with what it would be otherwise.

Figure 4.2 illustrates one form of Manchester encoding, which is applied to the data before it enters the modulator. At the top of the figure is a string of data represented in NRZ (nonreturn to zero) format, in which the data bits appear one after another, with nothing in between. A separate clock signal is typically used to transfer NRZ data from one place to another. However, in a communications link, a separate clock path does not exist. Therefore, in some manner the clock must be embedded with the data. Manchester encoding provides a simple, albeit bandwidth inefficient, way to do this.

The center portion of Figure 4.2 shows the bit cell boundaries: the time when the signal transitions from one bit to the next. At the bottom of the figure

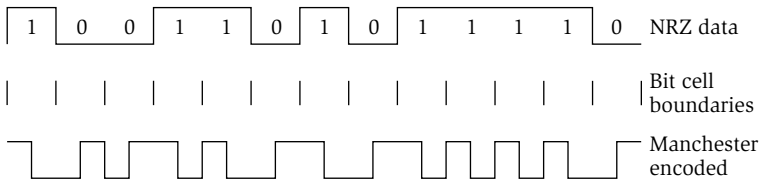


Figure 4.2 Manchester Encoded Data.

is shown the Manchester encoded data. In the middle of each bit cell, Manchester encoding causes the signal to transition from one level to the other. This regularly spaced set of transitions forms a clock for the receiver. The first half of each cell contains the value of the bit; the second half contains the complement (opposite state) of the bit (or vice versa). A variation is differential Manchester encoding: At the beginning of each cell containing a 0, the encoded signal makes a transition; at the beginning of each cell containing a 1, the transition is omitted. There is always a transition in the middle of the bit cell. On the other hand, if the transition occurs in the middle of a cell containing a 1 but not in a cell containing a 0, the encoding is called *biphase mark*. It is used in the S/P DIF interface described in Chapter 22.

Note that Manchester encoding is *not* a modulation method. Rather, it is a baseband coding method that is useful in low-cost systems to provide clock recovery and removal of a dc component of the data.

4.2.3 Baseband Spectrum of a Data Signal

The spectrum of a data signal may be understood with reference to the waveform of Figure 4.2 and the low-pass filtering of Figure 4.1. Figure 4.1 shows that it is possible (and usually desirable) to low-pass filter the data waveform so that when a string of alternating 1s and 0s (or a string of like bits in Manchester encoded data) is encountered, the waveform looks substantially like a sine wave. No useful data is lost in the filtering process, and the occupied bandwidth of the signal is thus minimized. Practical filters are somewhat wider than this, but the principle remains.

When a string of 1s and 0s appears in an NRZ signal (defined in Section 4.2.2.) and is optimally filtered, the resulting waveform is substantially a sine wave at a frequency of one-half the bit rate. To understand this, recall that a complete cycle of a sine wave occupies the time from one positive-going zero crossing to the next positive-going zero crossing. In order to produce such a waveform, two bits, 1 followed by 0, are required.

In general, a data signal can contain alternating 1s and 0s. This pattern defines the highest baseband frequency, equal to one-half the bit rate. However,

that data can also contain long strings of 1s or 0s, which represent a frequency approaching 0. Because of this, the fundamental spectrum of the baseband (unmodulated) data signal extends from one-half the bit rate down to very close to 0.

In summary, the spectrum of a digital signal extends from very low to at least one-half the bit rate (really the symbol rate, a concept to be introduced later in the chapter). Practical systems have frequencies that extend above this due to real-world filtering constraints. The excess bandwidth may range from a few percent to much higher, depending on the application.

4.2.4 Introduction to Bit Error Rate

An important figure of merit for any data communications system is the *bit error rate* (BER) of the system. This represents the proportion of the bits transmitted that are received incorrectly. If, on the average, two bits are in error for every million (10^6) bits transmitted, then the bit error rate is

$$\text{BER} = \frac{\text{number of bits received in error}}{\text{total number of bits transmitted}} = \frac{2}{10^6} = 2 \times 10^{-6} \qquad (4.1)$$

The BER is a function of the type of modulation, the characteristics of the modulator and demodulator, and the quality of the transmission channel. A common way to compare different modulation systems is to plot the BER against the signal-to-noise ratio (data communications practitioners tend not to differentiate between "signal" and "carrier" to noise). This is a convenient way to compare the capabilities of different modulation techniques and to predict performance in the presence of noise. This topic is covered in Section 4.2.12. Sometimes BER is plotted against signal-to-noise ratio in the presence of certain other impairments, such as group delay distortion, interference, and phase noise.

Obtaining an equation relating carrier-to-noise ratio to BER for an FSK system is difficult due to the nonlinear nature of the frequency modulation process. The literature contains only a few attempts, and contradictions exist. Within the parameters commonly used in cable television work (very wide deviation compared with the bit rate), the BER can be quite good at very low C/N. The penalty is that a significant amount of bandwidth is consumed, compared with that required for other modulation methods operating at the same bit rate.

4.2.5 Biphase Shift Keying (BPSK)

Biphase shift keying (BPSK) is the simplest of a class of modulation that is used extensively in cable television work. It is somewhat more complex to detect than is FSK, although producing a BPSK signal is simple. In a BPSK system, the phase is shifted 180° between transmission of a 0 and a 1.

Figure 4.3 illustrates a simple BPSK transmission system. An NRZ datastream is low-pass filtered (LPF1) to remove higher-order harmonics. The filtered signal is passed to a balanced amplitude modulator, M1, whose RF input is supplied from local oscillator LO1.

The figure illustrates the output of the modulator, which consists of a carrier envelope that drops to zero amplitude in the transitions between 0 and 1. The modulated carrier has one phase relation for a 0 and the opposite phase when a 1 is being transmitted. Note the phase of the carrier with respect to the carrier timing tick marks (every 180° of the carrier) shown below the RF envelope. The reduction in amplitude to zero between states is a direct result of the filtering action of LPF1. Were it not for this filtering, the transition between 1 and 0 would not require the amplitude of the carrier to drop, but the occupied bandwidth would theoretically be infinite.

At the receiver, the signal is bandpass filtered in BPF1. From the filter, the signal is supplied to mixer M2, the demodulator. The second input to M2 is from local oscillator LO2. This oscillator must be phase locked to the incoming RF carrier. A number of different carrier recovery circuits can be used. The circuit shown utilizes the fact that the second harmonic of the carrier contains no phase shift (when the frequency is doubled, a 180° phase shift becomes a 360° phase shift, that is, no phase shift at all). The incoming frequency is doubled by the times-two multiplier, ×2, and then used to lock LO2. The output of LO2 is divided by 2 and supplied as the second input to M2.

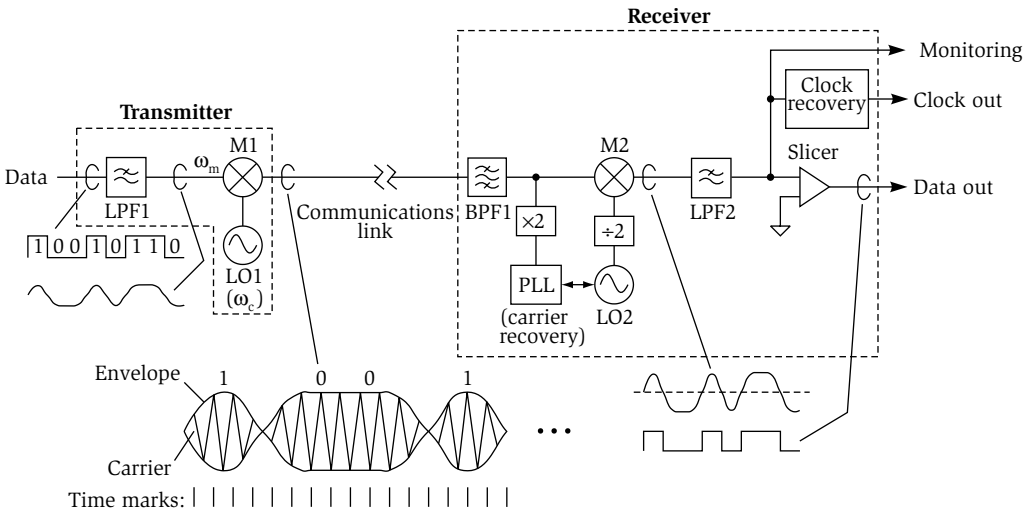


Figure 4.3 BPSK Transmission System.

Basic Math of BPSK

One might be tempted to look at the waveform of Figure 4.3 and conclude that BPSK is a phase modulation technique because the carrier reverses phase between transmission of 1s and 0s. Indeed, the name of the modulation leads to that conclusion. The correct way to visualize BPSK is as double sideband suppressed carrier amplitude modulation.

To understand the modulation process, assume that the data signal is an alternating series of 1s and 0s that has been filtered to a sine wave, following the logic of earlier sections of this chapter. Then the modulating signal can be expressed as

$$e_m = M \sin \omega_m t \tag{4.2}$$

where the subscript m indicates the modulating frequency and $\omega_m = 2\pi f_m$. As implied by Figure 4.3, the modulation process involves multiplying this modulating frequency by the carrier frequency, represented as

$$e_c = A \sin \omega_c t \tag{4.3}$$

The resulting modulated signal is expressed as follows, expanded through use of common trigonometric relations:

$$\begin{aligned} e_r &= e_c \cdot e_m \\ &= AM \sin \omega_c t \sin \omega_m t \\ &= \frac{AM}{2} \cos(\omega_c - \omega_m)t - \cos(\omega_c + \omega_m)t \end{aligned} \tag{4.4}$$

The center row shows the phase shift with modulation: When the sinusoid representing the data modulation ($\sin \omega_m t$) is positive, then the carrier has one phase; when it is negative, the carrier phase is reversed, as illustrated in Figure 4.3.

The bottom expression shows that the frequency components of the modulated signal exist at the carrier frequency ω_c minus the modulating frequency ω_m and at the sum frequency. Thus, the bandwidth occupied by the signal is twice the modulating frequency, as is the case with conventional double sideband amplitude modulation. A component at the carrier frequency does not exist.

The mathematics of the synchronous detector shown in Figure 4.3 consists of multiplying Equation (4.3) by the carrier sinusoid and applying more common trigonometric relationships:

$$
\begin{aligned}
e_d &= e_c \cdot e_r \\
&= \sin \omega_c t \cdot AM \sin \omega_c t \sin \omega_m t \\
&= AM \sin^2 \omega_c t \sin \omega_m t \\
&= \frac{AM}{2}(1 - \cos 2\omega_c t) \sin \omega_m t \\
&= \frac{AM}{2} \sin \omega_m t - \frac{AM}{2} \cos 2\omega_c t \sin \omega_m t
\end{aligned} \tag{4.5}
$$

The first term in the bottom line represents the recovered modulation; the second term represents components surrounding twice the carrier frequency. This latter term is filtered in LPF2 of Figure 4.3.

4.2.6 Quadrature Phase Shift Keying (QPSK)

BPSK modulation provides very good immunity to noise and is simple and low in cost. The problem is that it is inefficient in its use of the spectrum — better than FSK, but not good enough for high-speed data transmission. In order to make better use of the spectrum, it is necessary to take some small steps back toward (but not reaching) the analog domain, in that possible states are represented by different analog levels and/or phases. This allows the transmission of more than one bit at a time. The next modulation formats to be discussed are expansions of BPSK, which improve on the spectral efficiency by transmitting more than one bit at a time. The penalties are an increase in complexity and reduced immunity to noise and other transmission impairments.

Figure 4.4 illustrates a basic QPSK (also called 4-QAM for reasons that will become apparent) modulator and demodulator. Data entering the transmitter is split into two channels. For reasons that will soon be apparent, the two are called the I (in phase) and Q (quadrature) channels. Two bits are transmitted simultaneously, one in each channel. Spectral efficiency is improved because the spectrum required to transmit the two bits is no wider than is the spectrum required to transmit one bit using BPSK.

After the data is split into the I and Q channels, each channel is modulated onto a carrier, just as in BPSK. The same carrier frequency is used for each channel, but the phases of the two are 90° apart, or *in quadrature*. It can be shown that the two phases can be transmitted through a common communications channel, ideally without interfering with each other. (In the real world, there are several mechanisms that can cause the two channels to interfere, or *crosstalk*, but crosstalk can be controlled.)

After the two channels are modulated, the modulated carriers are combined for transmission to the receiver. At the receiver, a phase locked loop (PLL) locks a local oscillator, LO2, to the incoming carrier. The signal from LO2 is split into quadrature components. The in-phase component is sent to demodulator M3,

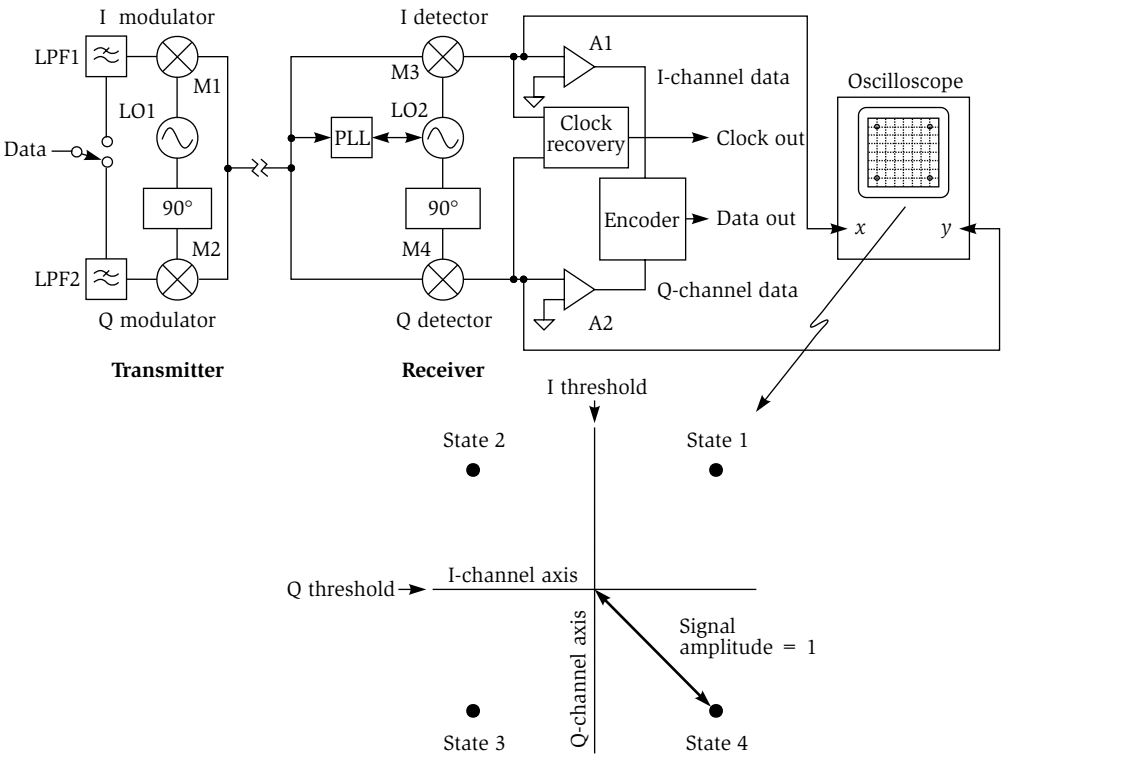


Figure 4.4 Basic QPSK Modulator and Demodulator.

and the quadrature component to the quadrature demodulator, M4. The two quadrature components of the local oscillator demodulate the corresponding component of the combined signal, while the opposite component is cancelled out because it is in quadrature phase relationship. An encoder combines the two bits from the I and Q channels into the resulting datastream. A clock recovery circuit reconstructs the clock signal from the incoming data.

The Constellation Diagram

A particularly useful representation of QPSK and higher-order signals is the constellation diagram, shown at the bottom of Figure 4.4. It is formed by routing the demodulated I signal, before slicing, to the *x* axis of an oscilloscope and the Q signal to the *y* axis. Each of the four possible positions of the symbol is formed from the I- and Q-channel data, with the position along the *x* axis being determined by the I data and that along the *y* axis by the Q data. The amplitude of the

signal is proportional to the length of the vector from the origin to any one of the four states, which are equidistant from the origin. Note that, in the absence of any transmission impairments, the four points in the constellation are small points. They are enlarged in the illustration to make them easier to see. As the signal transitions from one point in the constellation to another, the trajectory may pass through several other points.

The *y* axis forms the *decision threshold* for the I channel since it is halfway between the I channel states. Similarly, the decision threshold for the Q channel is defined by the *x* axis. The decision threshold defines the signal state at which the comparator(s) for that axis makes the decision that the bit represents one value or the other. If the point in the constellation crosses the decision threshold due to noise or other channel impairment, the comparator will yield the wrong output, resulting in an error in transmission. Computation of bit error rate (BER) is based on determining the probability that any one point in the constellation could be forced to cross a threshold into another state's place in the constellation. When this happens, at least one, and sometimes more, bits are received in error. Obviously, the closer together the states in the constellation (i.e., the "denser" the modulation), the less noise or other disturbance can be tolerated before a state can be misinterpreted as that of a neighbor.

The peak amplitude of the signal is equal to the length of a vector from the origin to the farthest point in the constellation. In the case of QPSK, all four points are equidistant from the origin, so any of them can be taken as the peak amplitude. Note that the constellation diagram represents voltage, not power.

Though the QPSK signal has four states at the same amplitude, between states the amplitude of the carrier drops, due to filtering at the modulator. This filtering is represented by LPF1 and LPF2 of Figure 4.4. Also see Figure 4.3 for an illustration of the resultant modulated waveform. Since QPSK is nothing more than two BPSK signals combined 90° out of phase, then the same waveform of Figure 4.3 would generally apply to a QPSK signal: Between symbols, the signal amplitude passes through 0 if the state of both bits in the symbol changes. This usually doesn't cause much concern in cable television work. However, in the case of systems having high AM/PM conversion, where a change in amplitude of the signal can cause a change in its phase, the amplitude of the signal dropping to 0 can be a problem. AM/PM (spoken as "AM to PM") conversion is a phenomenon in which changes in the amplitude of a signal result in a change to the phase of the signal. In data communications, it can distort the phase reference for the PLL in a receiver, making accurate recovery of the data less likely. AM/PM conversion can be caused by asymmetrical nonflatness of the channel and by capacitance modulation in an amplifier that is operating out of its linear range.

A satellite communications link is susceptible to AM/PM conversion. Often the power amplifier on the satellite is driven slightly into saturation and may exhibit AM/PM conversion. To reduce the effect, satellite communications often

uses *offset QPSK*, in which the transition times of the data in the I and Q channels is offset by one-half of a bit cell. Since the two channels don't change amplitude at the same time, the amplitude tends to be more constant.

4.2.7 Higher-Order QAM Modulation

Binary Numbering

Before we discuss higher-order QAM, we need to review a couple of elementary concepts in binary numbering. We will show groups of two bits whose value is designated as 00, 01, 10, or 11. The correct way to speak this is "zero-zero," "zero-one," and so on ("zero" may be replaced with "oh"). These numbers are not to be confused with the base-10 numbers that look the same. If confusion between base 2 (binary) and base 10 (our familiar numbering system) is possible, then the number will be followed by a subscript 2 or 10. The following are equivalent numbers in base 2 and base 10: $010_2 = 2_{10}$ and $1101_2 = 13_{10}$.

16-QAM Modulation

The next denser modulation format commonly found in cable television work is 16-QAM (16 state quadrature amplitude modulation). It is used for modem return services that require higher bandwidth efficiency than that offered by QPSK. It is similar to QPSK except that each axis is divided into four equidistant levels so that each channel (I and Q) can carry the information of two bits. Each of two phases carries two bits at a time, so there are four bits per symbol.

Figure 4.5 illustrates a basic 16-QAM modulator and demodulator. Note the similarities between this and the QPSK modulator of Figure 4.4. Each channel of the 16-QAM modulator can take on not only the two phases of QPSK but, in addition, intermediate amplitude values. Data is split into two channels, the I and Q channels. Instead of only one bit being routed to each channel at a time, two bits are routed to each. They are added as shown in the table beside the Q-channel modulator: for example, if the two bits are 11 (pronounced "one-one," since we are working in binary numbers), then the output of the adder is $+3$, and the Q-channel modulator puts out an amplitude of $+3$, with a positive phase (calling this the "positive" phase is arbitrary). If the data set is 10 (one-zero), then the adder supplies an analog signal of amplitude $+1$ to the modulator.

Similarly, a 01 data set yields an analog level of -1, which causes the modulator to supply a 1-level signal with the opposite phase. A 00 data set will cause the adder to supply -3 to the modulator, producing a higher-amplitude negative phase signal (saying the phase is "negative" is an arbitrary way to refer to the fact that the phase is opposite that of a "positive" phase). By convention, the inner levels are $+1$ and -1. To retain equidistance, spacing between levels is 2.

The same thing takes place in the I channel. Each channel can thus take on any of four states. With the I channel in any of four states and the IQ channel

independently in any of four states, a total of 16 states are possible. This is the number of states required to represent 4 bits ($2^4 = 16$). The modulation technique, 16-QAM, derives its name from the 16 possible states of the modulated signal.

The receiver contains in-phase and quadrature demodulators, as does the QPSK receiver. However, the output of each demodulator must be further processed to recover two bits. This is done for the I channel, using the three data slicers shown. The data slicers have thresholds set by the four voltage divider resistors, at levels of T1, T2, and T3. These levels are set halfway between the received states in the I channel. The thresholds are shown in the constellation diagram at the bottom of Figure 4.5. By reading the outputs of the three comparators, A1 through A3, the decoder can deduce the transmitted state. Note that there are more decision thresholds than there were in QPSK, because there are more points in the constellation, and we must have a way to distinguish between all states. We often refer to *decision boxes* rather than

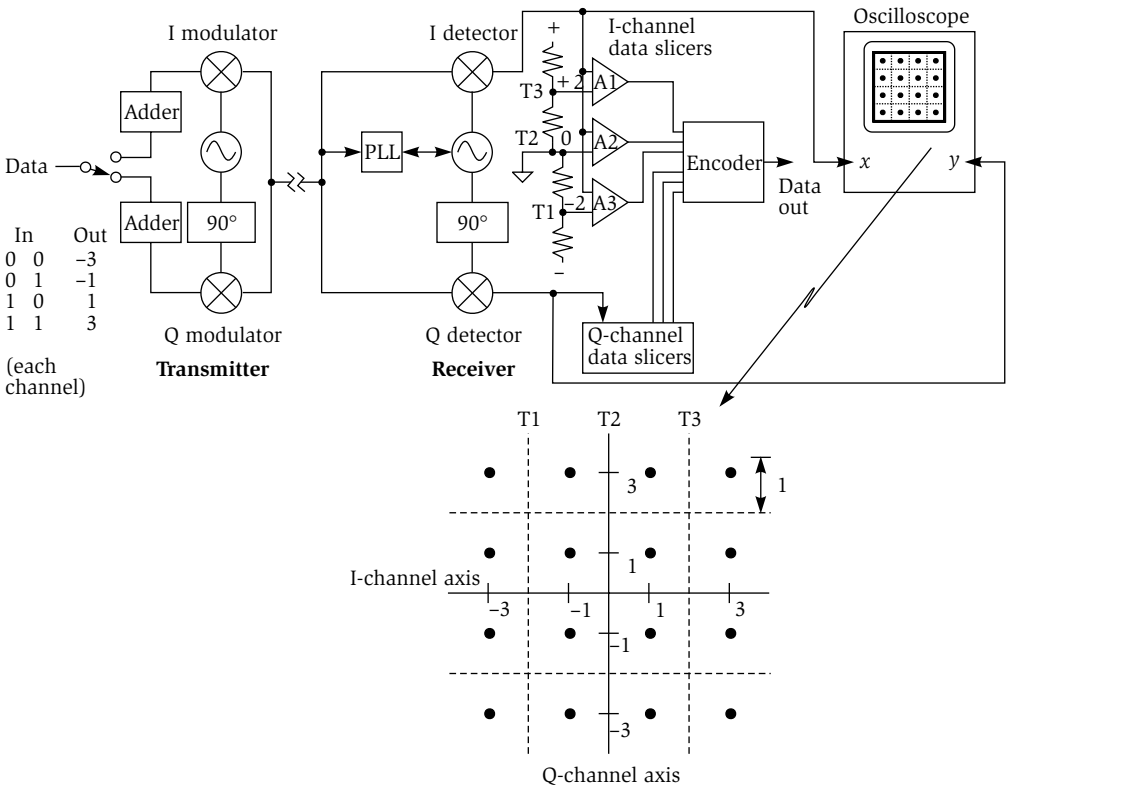


Figure 4.5 16-QAM Modulator and Demodulator.

thresholds, because each point in the constellation has its own box it must be in, defined by the I and Q decision thresholds.

64- and 256-QAM

For downstream transmission, the industry employs modulation formats that are similar to 16-QAM except that they have more states. They transmit more bits per symbol, so the constellation is denser. The distance between states is smaller, so the susceptibility to noise and transmission path distortion is greater. On the other hand, the spectral efficiency is greater, meaning that more bits can be transmitted in the same RF bandwidth.

Figure 4.6 illustrates the constellation diagrams for 4-QAM (i.e., QPSK), 16-QAM, 64-QAM, and 256-QAM. The diagrams are drawn to the same peak scale, so it is easy to see the higher density of the higher-order modulation formats. Also shown for comparison are several characteristics of the different

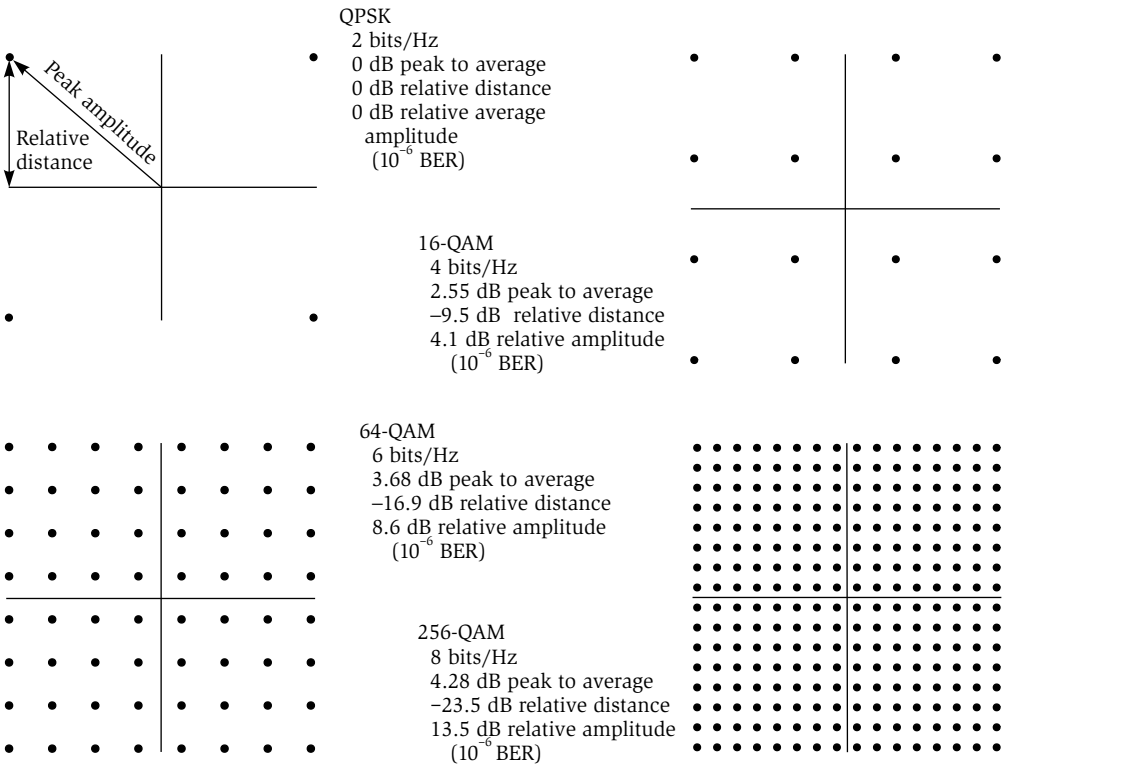


Figure 4.6 Constellation Diagrams, 4-, 16-, 64- and 256-QAM.

modulation formats. Most of the characteristics are shown measured in decibels with respect to a QPSK signal.

For each modulation type, the approximate power peak-to-average ratio is shown first. This is computed by averaging the power of each state, ignoring the effects of filtering. Later we show how this is done. A more precise peak-to-average ratio must include the effects of filtering. Next is shown the theoretical spectral efficiency, in bits per hertz. The spectral efficiency is a figure of merit, defining how many bits per second can be transmitted per hertz of RF bandwidth. The actual spectral efficiency will be less than this, due to the practical need for excess filter bandwidth.

Below the peak-to-average ratio is shown the relative distance between a state and the closest state boundary, expressed in decibels. The relative distance between a state and the boundary is defined in the QPSK constellation, as is the peak amplitude. It is the decibel measurement of how much closer to each other the states are in the denser modulation formats. The reference is a QPSK constellation, and the distance is normalized to the peak, not the average, amplitude.

The relative amplitude entry is derived from the BER vs. signal-to-noise ratio of the equation graphed in upcoming Figure 4.17. It is an indication of the increase in signal-to-noise ratio required to maintain a 10^{-6} bit error rate. Recall that the signal level to which we refer here is the average, not the peak level.

32-QAM Modulation

A similar modulation format to those shown earlier is 32-QAM, which has 32 states and a theoretical spectral efficiency of 5 bits per hertz. Two different 32-QAM constellations are shown in Figure 4.7. Figure 4.7(a) is distinguished from the 64-QAM formats shown earlier in that every second state is occupied. This puts more distance between states, resulting in easier decoding. Figure 4.7(b) illustrates an alternate encoding of 32-QAM, where the four corner states are omitted. Both encoding methodologies have been reported in the literature.

A total of 32 states represents 5 bits ($2^5 = 32$). With 16-, 64-, and 256-QAM, each axis, I or Q, can represent a certain number of bits, because the number of states along each axis is a power of 2. On the other hand, 32-QAM does not have this correspondence between bits and states possible. The five bits making up a symbol can, however, be encoded into one of the 32 possible states.

4.2.8 8-PSK Modulation

Because satellite transponders normally operate close to or in compression, the modulation transmitted over a satellite path must not depend on amplitude discrimination to recover data. For many years, QPSK, or 4-QAM, has been the digital modulation of choice for satellite transmission. Recently 8-state phase shift keying (8-PSK) has begun to see use in order to improve bandwidth efficiency in satellite communications. Figure 4.8 illustrates the constellation

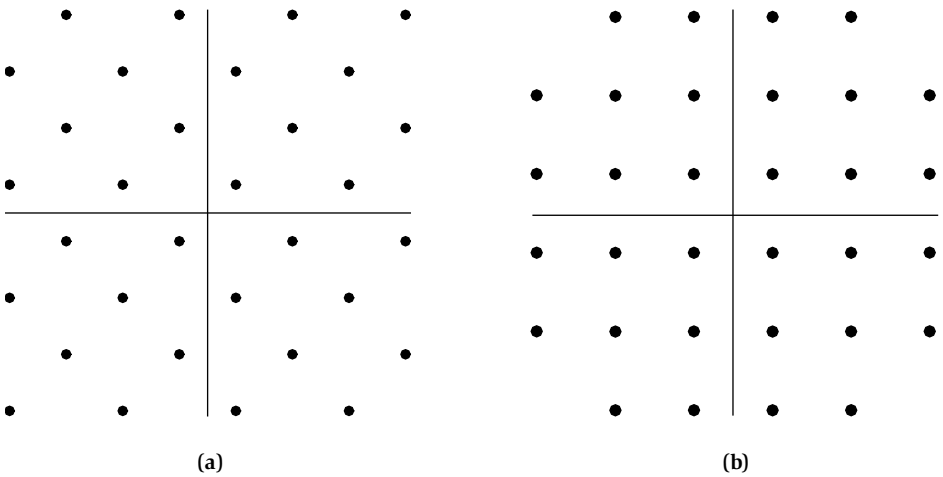


Figure 4.7 32-QAM Constellations. (a) Alternate states. (b) Corners removed.

of an 8-PSK modulated signal. It has eight states (3 bits per symbol); but since they are all equidistant from the origin, all states have the same amplitude (there will be some amplitude reduction between states due to filtering). Thus, the satellite amplifier operates at the same output amplitude at all times, a necessary condition for efficient amplifier operation.

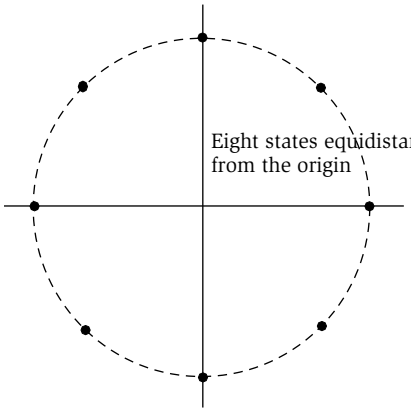


Figure 4.8 8-PSK constellation.

4.2.9 8-VSB Modulation

The FCC has designated 8-VSB modulation as the standard for off-air television in the United States. It is very similar in performance to 64-QAM. In VSB modulation formats, only one phase of the carrier is used, and there is no quadrature component; all bits making up the symbol modulate the same phase of the carrier. That carrier (which is almost completely suppressed) is placed very near the band edge rather than in the center, and one sideband is removed by filtering.

Both 8-VSB and 64-QAM have essentially the same spectral efficiency. 64-QAM carries data on two quadrature phases of the carrier but must place the carrier in the center of the passband. Where carriers in quadrature are employed, it is necessary to transmit both sidebands to allow for separation of the two channels. 8-VSB only uses one phase of the carrier, so separation of components is not an issue. Compensating for the lack of the quadrature phase, 8-VSB is able to transmit twice the symbol rate in the same bandwidth, since the second sideband is not transmitted.

Figure 4.9 illustrates the principle of VSB digital modulation. A set of bits is converted to a parallel word. The number of bits taken at a time is three when 8-VSB ($2^3 = 8$) is being transmitted. These bits are added, and the result is converted to analog and supplied to a balanced modulator. The modulator output is filtered in a vestigial sideband (VSB) filter. Ideally, the VSB filter doesn't pass any of the lower sideband but does pass the carrier. As a practical matter, some of the lower sideband must be transmitted, as in the case of analog transmission. Advances in filter technology since the NTSC television system was developed have made a much smaller vestigial sideband possible. In analog VSB transmission, the picture carrier is 1.25 MHz above the channel edge. In digital VSB transmission, the carrier is only 310 kHz above the channel edge.

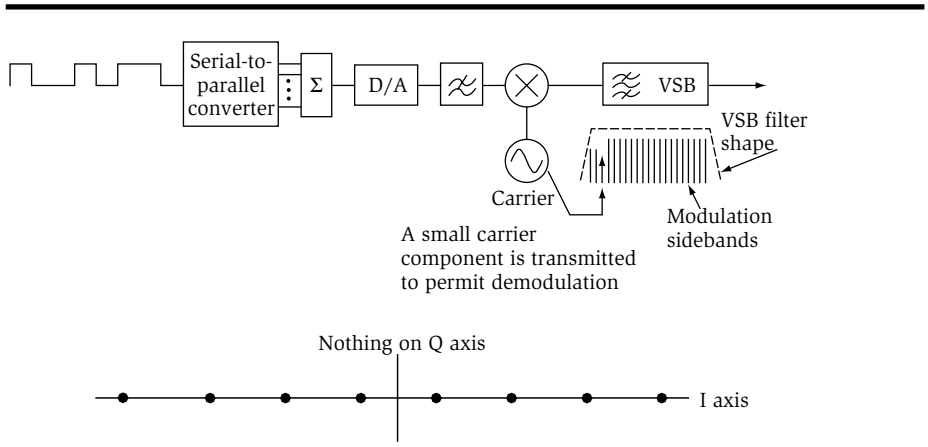


Figure 4.9 VSB Modulator.

A small amount of carrier is inserted in the VSB signal to facilitate carrier recovery. With QAM modulation formats, the carrier may be recovered by taking advantage of the symmetrical sideband nature of the signal. It may be shown that a suppressed carrier can be recovered by doubling the frequency of the received signal, which yields a second harmonic of the carrier. This is true only when the carrier has symmetrical sidebands. Such is not the case with VSB, so a small carrier component must be transmitted to allow recovery.

The susceptibility to transmission impairments of an 8-VSB signal is similar to that of a 64-QAM signal. VSB has a claimed advantage in the presence of phase noise. As shown in upcoming Figure 4.12, phase noise causes the constellation to rotate around the origin. With VSB transmission, the distance to the state boundary on the phase trajectory is much greater than it is with QAM. The excess bandwidth of the broadcast 8-VSB signal is 11.5%. The symbol rate is 10.76 Mbaud, so with three bits per symbol the gross data rate transmitted is 32.28 Mb/s. Due to trellis coding, there are only two information bits per symbol, so the gross payload is two-thirds of this, or 21.52 Mb/s. Adjusting for overhead, the net transport stream payload seen by a transport decoder is 19.393 Mb/s.[1] Sometimes you may see the data rate shown as 19.29 Mb/s. This number is the data rate transmitted (less modulation overhead), where the MPEG sync bit is not transmitted but rather is added back at the demodulator. This data rate (whichever number you use) is less than the 64-QAM data rate transmitted in cable, primarily to allow for more error correction to compensate for the perils of the radio path.

4.2.10 Fundamental Principles of Digital RF Transmission

In this section we describe a number of issues associated with digital modulation. You should be familiar with them in order to make sense of what you experience related to a digitally modulated signal.

Bit Rate and Baud Rate

Perhaps the most common mistake made in communications terminology today is to confuse the rate at which *data* is transmitted with the rate at which *symbols* are transmitted. The rate at which bits are transmitted (the *bit rate*) is measured in *bits per second*. In most modulation formats, however, more than one bit is transmitted at a time. The bits that are transmitted simultaneously are grouped into a *symbol*. The rate at which symbols are transmitted is the *baud rate*.[2]

In DOCSIS upstream transmission, the lowest specified symbol rate is 160 kbaud. If QPSK modulation is used, the bit rate is twice that, or 320 kb/s, because there are two bits transmitted per symbol. However, if 16-QAM is used, the bit rate is four times the baud rate, or 640 kb/s. (Note that the plural of *baud* is *baud*, without an "s.") As shown earlier, 256-QAM modulation groups 8 bits into a symbol, so the baud rate is one-eighth of the bit rate. The

occupied bandwidth of the signal is a function of the baud, or symbol, rate, because the baud rate tells us how many times the modulation state changes each second.

Filtering Methodology

This section introduces the filtering methodology that is followed in many data communications applications and for many types of digital modulation (except FSK, which was covered earlier). The filters at the transmitter and receiver are effectively in cascade in the signal path; together it is desirable that they form a filter just wide enough to convert the highest modulating frequency into a sine wave. (As a practical matter, the filter will be somewhat wider than this.) A particularly useful class of filters is called *raised cosine* filters by the data communications industry. They are called *sine squared* filters by the television engineering community.

The equivalence is seen in the following trigonometric relationship:

$$2\sin^2\alpha = 1 - \cos 2\alpha \tag{4.6}$$

The left side of the equation shows the rationale behind the television usage of *sine squared* terminology, while the right side shows the basis of the data communications term *raised cosine*. The 1 on the right "raises" the cosine term, which by itself ranges from -1 to $+1$. In television work, a sine squared filter is used to produce the well-known sine squared pulse and bar, which are useful in analyzing analog video channel performance. In data communications, the cascaded performance of the low-pass filters at the transmitter and receiver, possibly in addition to the response of the IF filters, form a raised cosine filter, which ensures a fast transition between the modulation states, consistent with minimal spectrum occupancy.

It is desirable to heavily filter at the transmit end in order to reduce the spectrum usage, and heavy filtering is desirable at the receive end to minimize noise added during transmission. However, the total filtering applied must conform to the raised cosine shape. In many communications systems, the designer places one-half of the filter function at each end of the communication channel. When this is done, the filter response at each end is the square root of the complete filtering function, so the filter used on each end is often described as a *root raised cosine* filter. Note that the filters at the transmit and receive ends must be matched to each other and collectively are often referred to as *matched filters*.

Note that any bandpass filter must be symmetrical with respect to the center (carrier) frequency for quadrature signals (that is, any QAM signal). This is a requirement of quadrature transmission systems. If the overall filter response is not symmetrical, we can show that the carrier phase will be recovered incorrectly, resulting in a loss of signal integrity. In addition, the in-phase and quadrature channels will crosstalk.

Required Bandwidth of the Signal

We shall develop the concept of occupied bandwidth using a BPSK signal, and then show how it is extended to any QAM modulation format. The bandwidth requirements of BPSK can be understood by considering the worst case, that of an alternating string of 1s and 0s. This will produce a square wave, one cycle of which is composed of an alternate 1 and 0, or two bits, as explained earlier. Thus, the frequency of the square wave, f_m, is one-half the bit rate. Figure 4.10 illustrates the spectrum of the data signal at baseband. The square wave of alternating 1s and 0s illustrates the highest frequency of the data signal. From Fourier analysis, we know that a square wave consists of all odd harmonics of the fundamental frequency, as shown by the vertical lines in the spectrum plot of the figure.

In real data systems, the data is not an alternating series of 1s and 0s. Rather, it is all possible combinations of strings of 1s and 0s. The resulting spectrum will fill in the spectrum from the maximum frequency, f_m, down to close to 0 frequency (and all harmonics thereof).

However, the harmonics are not needed to convey the information contained in the datastream. If all frequencies above f_m are removed by filtering, then an alternating series of 1s and 0s will produce a sine wave of frequency f_m.

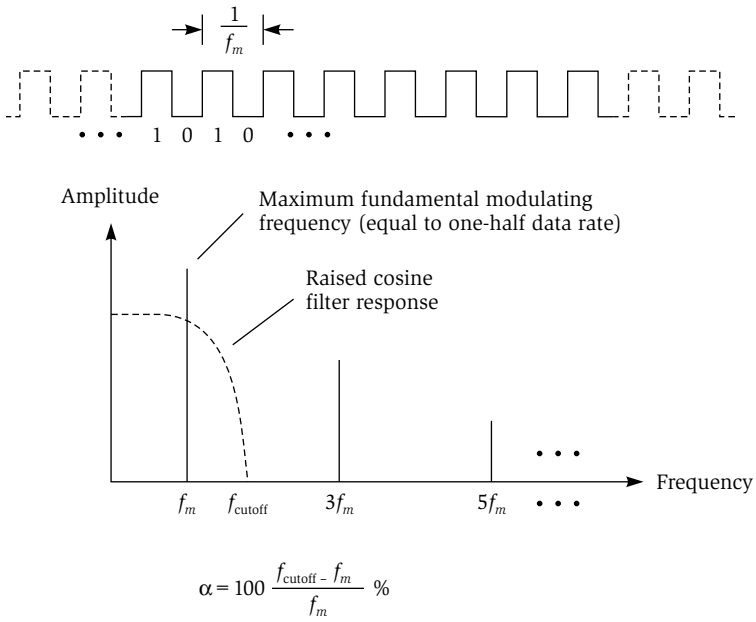


Figure 4.10 Baseband Spectrum of BPSK Signal.

Transitions that occurred at a slower rate (i.e., from a series of like states) would follow a sinusoidal transition path but would be flattened on top.

The *theoretical* (not practical) spectral efficiency of a BPSK signal is developed as follows. The alternating 1–0 sequence defines the highest frequency of the data baseband signal. If it is filtered by removing all harmonics, leaving a sine wave, then the frequency of the sinusoid is one-half the data rate (one cycle is composed of a set of a single 1 followed by a single 0). The bandwidth of the modulated signal is twice the modulating frequency (the sum less the difference between the carrier and the modulating frequency), which in turn is equal to one-half the bit rate. Thus, the spectral efficiency is two times one-half the data rate, or one times the bit rate, one bit per hertz.

A filter that completely cuts off just above f_m may be ideal but is hardly practical. A practical filter has the raised cosine shape and is somewhat wider than f_m. The sharper the filter cutoff (that is, the less bandwidth required between the filter passband and the stopband), the less bandwidth is required. However, the sharper filter is more difficult to build, and more group delay (see Glossary) is likely, depending on the realization. The shape of a real filter is described by the *alpha* of the filter, where alpha is defined as the excess of occupied bandwidth compared with the minimum required bandwidth:

$$\alpha = 100\frac{f_{\text{cutoff}} - f_m}{f_m}\,\% \tag{4.7}$$

Recall that the common filtering practice is to put the square root of the raised cosine filter response at the transmitter, and the same at the receiver. For example, as shown in Chapter 5, the filter specification for the transmitter of a DOCSIS-compliant downstream modulator operating in the 64-QAM mode is an 18% root raised cosine filter, meaning that the filter is 18% wider than the minimum required bandwidth.[3] This may also be interpreted that the occupied bandwidth is 18% wider than the theoretical minimum.

QPSK works the same way as does BPSK as far as the spectrum is concerned. QPSK is simply two BPSK carriers at the same frequency, but 90° out of phase with each other. The spectra overlap; and since a spectrum analyzer doesn't respond to the phase of a signal, the BPSK and QPSK spectra are very hard to tell apart.

Higher orders of QAM modulation are simply the same thing as QPSK, with the exception that rather than each of the two quadrature carriers taking on only two states (phases), they can take on multiple amplitude and phase states. The spectrum is again very difficult to distinguish from that of a BPSK signal. The bandwidth is based on the number of symbols transmitted per second, not on the number of bits transmitted per second. See the earlier discussion of bit rate vs. baud rate.

Clock Recovery

While self-clocking formats, such as Manchester encoding (see Section 4.2.2), could be used with BPSK and higher-order modulation, this is normally not done. The higher number of transitions associated with Manchester encoding would double the required bandwidth of the signal. Rather, an NRZ datastream is transmitted. Clock recovery consists of synchronizing a clock to the transitions in the received data. Where no transition occurs, the clock "freewheels" through the place where the transition should occur. If a long transmission of 0s or 1s occurs, it is possible to lose clock synchronization, so "scrambling" or some other technique is often used to force a state change every so often.

Data Scrambling

Scrambling is often used as a method of dc removal, which is necessary to effect proper operation of a number of circuits in the receiver, including clock recovery. *Scrambling* in its data transmission sense is not the same as when used in the cable television sense. In cable television, the word refers to the process of rendering a video and/or audio signal unintelligible to a normal receiver. A special circuit must be used to restore the signal to its intelligible form.

In data transmission, scrambling is the process of combining a datastream with a pseudorandom bit sequence, with the intent to remove any possible dc component of the datastream. An equivalent way of expressing the function of the scrambler is to say that it equalizes the power density over the occupied bandwidth. It can be shown that, if any data signal is exclusive ORed with a pseudorandom signal having a flat spectrum and no dc component, then the resultant signal will have a flat spectrum and no dc component. The signal can be recovered at the receiver by exclusive ORing the signal with the same pseudorandom signal.

The exclusive OR operation is a logical operation on a single bit of data. If the two bits are the same, the output of the exclusive OR operation is a logical 0; if the two bits are different, the output is a logical 1.

Differential Encoding

The phase locked loop used for carrier recovery in Figure 4.3 exhibits a problem, in that a phase error ambiguity exists: If the output of LO2 divided by 2 is in phase with the carrier applied to M1, then the recovered data will have one polarity; if it is out of phase, the recovered data will have the opposite polarity. Since the loop is closed at twice the frequency of the carrier, no simple way to determine the correct phase exists. One way would be to detect a code word buried in the data and to correct for a phase error.

A method frequently used to resolve the phase ambiguity is *differential encoding*. In differential BPSK, a 0 is represented by no change in phase from

the previous state; a 1 is represented as a 180° phase change. The detector is arranged to compare the phase of the previous bit with the phase of the present bit; if the phase doesn't change, a 0 is delivered as the received bit. If the phase is changed, a 1 is delivered (or vice versa). With more complex modulation schemes, sometimes a recognizable binary word is used to determine the correct phase.

The Eye Diagram

The recovered waveform at the "monitoring" point shown in Figure 4.3 should be the same as the waveform applied to modulator M1. However, distortion in the transmission path will change the wave shape, as suggested in Figure 4.11 (the dotted line is the original wave shape; the solid line is the received wave shape). Noise will put "ripple" on the waveform, as seen on an oscilloscope. If enough noise is added, then the state of the signal may be misinterpreted, producing a bit error. Besides noise, group delay can distort the wave shape, as can echoes and amplitude tilt in the channel.

A particularly useful oscilloscope display is called an *eye diagram*, so named for its resemblance to an eye. It is produced by supplying the received

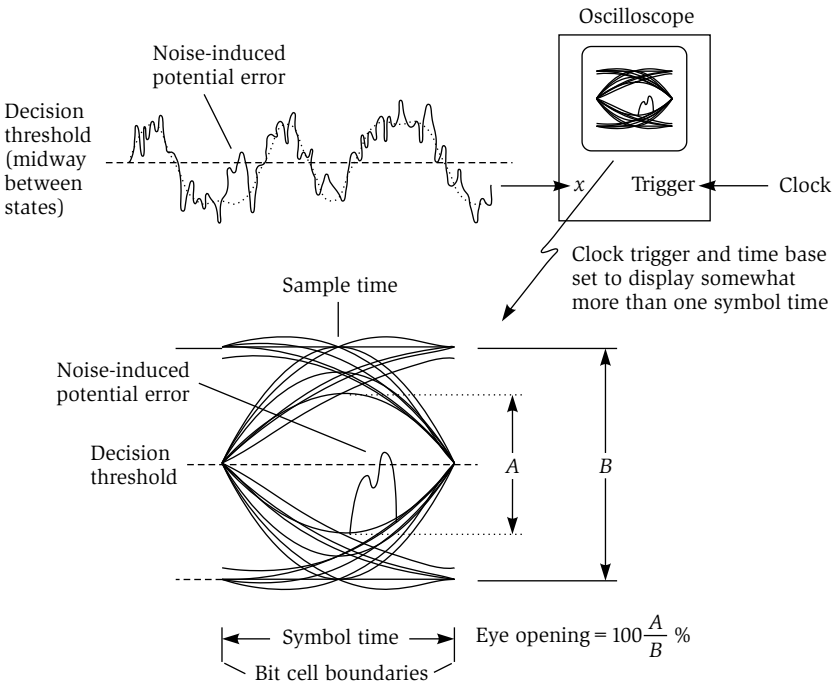


Figure 4.11 BPSK Received Waveform and Eye Diagram.

waveform (before the slicer) to the vertical input of the oscilloscope. The horizontal time base is set to display more than one bit cell and is triggered by the recovered clock. (For simplicity, the diagram shows only one bit cell width, though typically some additional time would be displayed.) When a bit is undergoing a transition from 0 to 1 at the trigger time, the waveform will be positive-going on the left. When it is in transition from 1 to 0, it is negative-going. It may also be the same as the last bit, in which case it forms a flat line at the top or bottom of the display. Similarly, the display toward the right is formed by transitions to the next bit after one displayed.

The slicer interprets the signal to be either a 1 or a 0. The sample time should be in the center of the bit cell, when the eye is as "open" as possible. One of the practical error sources, which is often not considered in computing BER versus noise, is errors in the timing of the sample: If the timing is in error because of problems in the clock recovery circuit (perhaps due to noise, interference, or waveform distortion), then it is more probable that the bit will be misinterpreted. A figure of merit is the *eye opening*, or the percentage of the peak (fully opened) eye that is seen as open in the eye diagram.

Eye closure can be caused by a number of transmission problems. The effect of noise is illustrated in Figure 4.11. Besides noise, errors in the filter shape can cause pulse distortion, which will tend to close the eye. If the transmission path exhibits excessive group delay or if an echo exists, then one bit may not completely "go away" before the next bit starts. This condition is called *intersymbol interference*, or ISI.

Illustrated in Figure 4.11 is a potential bit error, in which a noise spike carries the waveform so far from its correct position that it actually crosses the decision threshold of the slicer. If the clock causes the slicer output to be read at the time of the noise spike, then a bit error would occur.

We have illustrated the eye diagram for a BPSK signal. When quadrature modulation is used, such as in 4-, 16-, 64-, or 256-QAM modulation, the eye diagram is displayed separately for each axis of modulation. When there are more than two states per symbol, then the eye diagram will be more complex but will nonetheless resemble that of Figure 4.11.

Effects of Interference on the Constellation Diagram

The constellation diagram is useful for understanding the effect of interference on the signal. In Figure 4.12, quadrant I illustrates the effect of a continuous wave (CW) interfering signal in the passband of the receiver. An interfering carrier creates a beat note, which forces each point in the constellation to spread out in a circle around the desired point. The radius of the circle is proportional to the relative amplitude of the interfering carrier. Unlike analog video signals, the impact of the interfering carrier tends to be independent of its location within the passband, except for some effect caused by nonflatness of the receive filter.

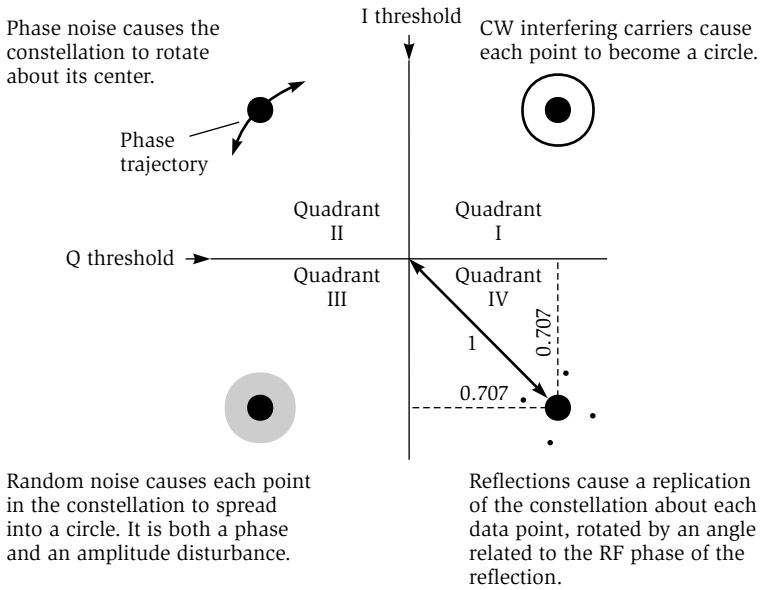

Phase noise causes the constellation to rotate about its center.
I threshold
CW interfering carriers cause each point to become a circle.
Phase trajectory
Quadrant II
Quadrant I
Q threshold
Quadrant III
Quadrant IV
1
0.707
0.707
Random noise causes each point in the constellation to spread into a circle. It is both a phase and an amplitude disturbance.
Reflections cause a replication of the constellation about each data point, rotated by an angle related to the RF phase of the reflection.

Figure 4.12 The Effect of Interference on the Constellation Diagram.

It is worth noting that composite second order (CSO) and composite triple beat (CTB) are not CW interferences. Their amplitudes change constantly, with peaks reaching 15 dB above the average. The variation is due to phase alignment of the multiplicity of carriers. The occurrence of a peak is rare, but the peak can last from a microsecond to a few hundreds of microseconds (longer than the burst protection period of the Interleave.[4]

Quadrant II illustrates the effect of phase noise on the signal. Phase noise is introduced when a local oscillator has significant phase noise. It can also be shown to occur when the transmission channel does not have a response that is symmetrical about the carrier. Phase noise causes the constellation to rotate about the origin.

Random noise (quadrant III) causes each point in the constellation to "blur," somewhat as with CW interference, except that the spreading of the point tends to be more nearly uniform within the circle, and the circle does not have a well-defined radius. Again, recall that the spreading of points in the constellation becomes a problem when the spread points cross a threshold to another state. With QPSK, that threshold is defined by the axes of the constellation diagram. With more dense modulation, the thresholds are closer together, so lower levels of interference will cause the threshold to be crossed. As a practical matter, an error may occur when a point is forced close to a threshold boundary: Real demodulators don't work quite as well as the theory predicts.

Finally, reflections, which cause ghosting in analog transmission, cause a replication of the constellation around each point in the constellation, as illustrated in quadrant IV. The constellation is rotated with respect to the "direct path" constellation, due to phase shift between the direct and reflected RF carriers. Usually, systems that employ denser modulation formats include adaptive equalizers, which can compensate for echoes if the echo doesn't change too quickly.

If the peak amplitude of the signal is 1, as shown in quadrant IV, then the distance from the nominal location of any of the four states to the nearest state boundary is 0.707. This result is easily obtained via simple geometry, where the amplitude vector, of length 1, is the hypotenuse of a right triangle and the distance to the nearest decision threshold is one of the two other sides of the triangle. Bear in mind that it is more common to measure the amplitude of a digital signal by using the average amplitude, which concept is explored later.

Effect of Gain Compression on the Constellation Diagram

Gain compression can occur in any amplifier stage, but it is most likely to be recognized as such if the amplifier is handling only a single digitally modulated signal. Gain compression may be described as a voltage transfer function having second- and third-order components, just as when we are describing distortion in broadband amplifiers. As the input signal instantaneous amplitude increases, the output amplitude increases more slowly. If an amplifier is handling multiple signals, this phenomenon produces composite second- and third-order distortion.

Figure 4.13 illustrates the effect of gain compression on the digitally modulated signal. The uncompressed constellation points are represented as dots, and the compressed constellation points are represented as pluses. In both cases we have normalized the signal amplitude to the same average power. This represents what a receiver would do if its automatic gain control (AGC) responded to average power. The decision thresholds are at the same point for both the uncompressed and the compressed constellations, due to the normalization. Note that the uncompressed constellation points are in the center of the decision boxes, as far from the decision boundaries as possible, where they should be for best data recovery.

The compressed constellation has very many points near the edge of the decision boundary; a slight amount of noise or other impairment would drive the constellation point outside of the boundary, forcing a bad decision on the transmitted information. For illustration, we have chosen a rather large amount of gain compression. But you can see that the problem gain compression causes is that the constellation is moved from the optimum location, with the result that more errors will be made for a given impairment level.

Peak-to-Average Ratio

Higher-order modulation techniques exhibit a higher peak level than the average level read on a power meter. Figure 4.14 illustrates one reason for this, based on

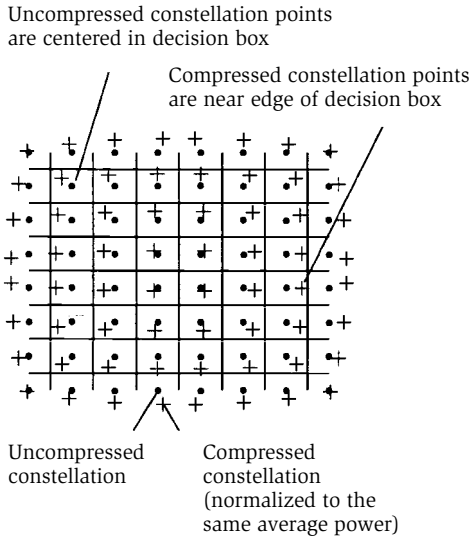
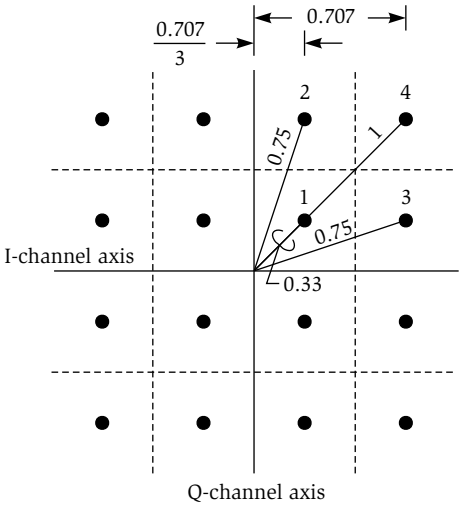


Figure 4.13 Effect of gain compression on a 64-QAM signal.



Element	Length (voltage)	Length^2 (power)
1	0.33	0.11
2	0.75	0.56
3	0.75	0.56
4	1.00	1.00
Average:	0.71	0.56

Peak to average: $20 \log \frac{1}{.71} = 3$ dB $\quad 10 \log \frac{1}{.56} = 2.55$ dB

Figure 4.14 Peak and Average Level of a 16-QAM Signal.

a 16-QAM signal. It shows the same constellation diagram as depicted in Figure 4.5. The constellation is scaled such that the amplitude of any one of the four corner states is 1. This represents the peak level of the signal. The chart in the figure computes the average signal level based on equal occupancy of any of the 16 possible states. Since all four quadrants behave the same way, it is necessary to average over only one quadrant.

The voltage amplitude of each state is determined by simple geometry, assuming that the highest-amplitude state has an amplitude of 1. From the geometry of a right triangle, it then follows then that the coordinates of that state are (0.707, 0.707). The state closest to the center has a distance on each axis of one-third of this, so its total distance is one-third that of the farthest state. States 2 and 3 each have one coordinate of 0.707 and one coordinate that is one-third of that. Application of the Pythagorean theorem shows that the magnitude is

$$\sqrt{0.707^2 + (0.707/3)^2} = 0.75$$

When looked at on a voltage basis, the voltage average is the average length of the four vectors that define the states in the quadrant. The power average is obtained by squaring the voltage vector for each state, because power is equal to the square of the voltage divided by the resistance:

$$P = \frac{E^2}{R}$$

The resistance would cancel out in the computation, so it is omitted. The right-hand column in the chart in Figure 4.14 computes the power peak-to-average ratio, based on squaring the voltage vectors.

Effects of Fitering on Peak-to-Average Ratio

The peak-to-average ratio is not completely characterized by the state distribution as implied, however. It is also a function of the filtering at the transmitter. This was illustrated in Figure 4.3, which shows an amplitude dip in a BPSK signal when the transmitted bit changes states. The same thing happens with higher-order modulation techniques, and that somewhat changes the peak-to-average ratio. The peak-to-average ratio is of concern in cable systems, particularly in the operation of return plant, where a limited number of digital signals comprises the return spectrum handled by the return laser.

It was shown that the peak-to-average ratio is partially a function of the relative amplitude of each state and the percentage of time the signal spends in each of those states. Another factor, not yet taken into account, is the filtering applied to the signal. As the filtering becomes more aggressive (i.e., lower alpha, corresponding to smaller excess RF bandwidth), the filter tends to ring more,

meaning that the signal will be of even greater amplitude than predicted, for a small percentage of the time.

Ito[5] studied this theoretically for a filter having an alpha of 11.52%. He computed the percentage of time the power of a 64-QAM signal exceeded its average by various amounts. The curve follows the percentage computed based on state occupancy, until the percentage of time is less than the percentage of time the signal spends in the highest state. He found that the signal exceeded the average by 6 dB or more for about 0.1% of the time and was 9 dB above the average for about 0.00001% of the time. (For reference, 6.25% of the time a 64-QAM signal is in its highest-amplitude state, which is 3.68 dB higher than the average power.)

Multiple-carrier systems, such as OFDM (described in Section 4.3.3), had a much higher peak-to-average ratio. It exceeded the average by 12 dB for 0.00001% of the time. This is because multiple-carrier modulation signals tend to have the amplitude distribution of random noise, which has a finite probability of being very high for a very short time. The subject of peak-to-average ratio is revisited in Chapter 16.

Effect of Linear Distortion on Pulse Shape

Linear distortions that can enter the cable plant include *echoes*, also called *reflections*, in which a pulse is reflected from an impedance mismatch and adds constructively or destructively to the original waveform. Another linear distortion is *group delay*, in which the delay of the channel is not flat with frequency. Yet another linear distortion is *amplitude ripple*, in which the frequency response of the channel is not flat. All of these linear distortions are familiar to those who have dealt with them in analog modulation. In digital modulation, these distortions tend to close the eye of the data (illustrated in Figure 4.11) and can present a serious impediment to data recovery.

Figure 4.15 illustrates pulse distortion caused by a single reflection. In the modeling, the reflection was delayed about 75% of the pulse duration. In many cases the pulse would be delayed much more. Also, to make it easier to see the effect of the delayed pulse, we have assumed a much wider excess bandwidth than is normal. You can see that the pulse is significantly distorted, which would result in significant closure of the eye diagram. Since the delay in the reflected signal may be much more than one symbol time, we often refer to these distortions as producing *intersymbol interference*, or ISI.

Adaptive Equalization

Because ISI can be so serious a problem in dense modulation formats, it is desirable to compensate for it at the receiver. This is done with an *adaptive equalizer*, a circuit that produces an echo that is the same as the echo added by the system, except the intentionally produced echo is subtracted from the signal.

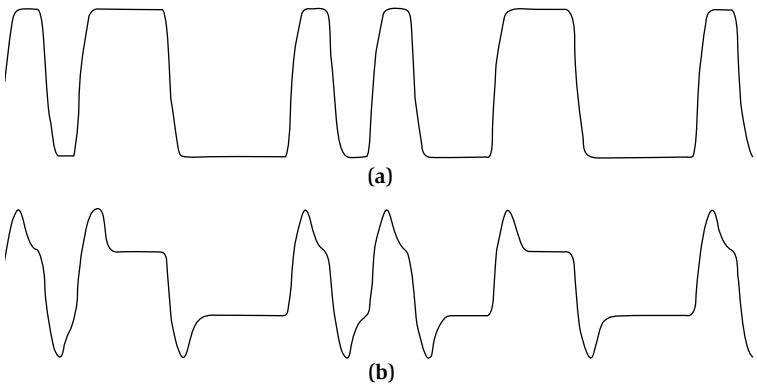


Figure 4.15 Pulse distortion caused by a reflection. (a) Original pulse. (b) Distorted pulse.

The higher the density of the modulation (e.g., 256-QAM), the more important the adaptive equalizer is, because less eye closure can be tolerated before bit errors occur.

Figure 4.16 illustrates an adaptive equalizer installed in each channel of a QAM demodulator. Figure 4.16(a) shows a QAM demodulator, with an adaptive equalizer installed in each channel, after the demodulator and before the data slicers. Thus, we have an analog signal at the point where the adaptive equalizers work. (Note that in modern circuitry, the signal is often digitized before detection, and the detection, filtering, and equalization are performed using digital signal-processing techniques. The point is that the signal at the adaptive equalizer is a continuous signal, even if it is represented digitally. It has not been reduced to the final binary value, which is done in the data slicers and the encoder.)

Figure 4.16(b) shows a 24-stage adaptive equalizer. The signal from the demodulator enters at the left and is successively delayed by small intervals (usually one symbol in time), designated Z^{-1}. After each delay, the signal amplitude is adjusted by complex gains A1, A2, etc. The signal with gain adjusted is added to the main signal, which is shown coming through A1 before any delay, plus the output of all other stages of the adaptive equalizer. Note that the gains A may be negative values, to subtract signal if necessary; depending on the RF phase of an echo, the reflected signal may add, subtract, or have any phase angle in between.

Thus, the adaptive equalizer generates an echo similar to that introduced in the plant and then subtracts it from the main signal in order to cancel the echo. The small eye diagrams at the left and right illustrate the effect of the adaptive equalizer (for simplicity, we are showing a BPSK eye, but the adaptive equalizer is more likely to be needed for dense modulation). Of course, it is necessary to adjust the value of each "tap" (each gain A) to compensate for whatever deficiency exists. The gains A may be positive or negative. It is the job of the

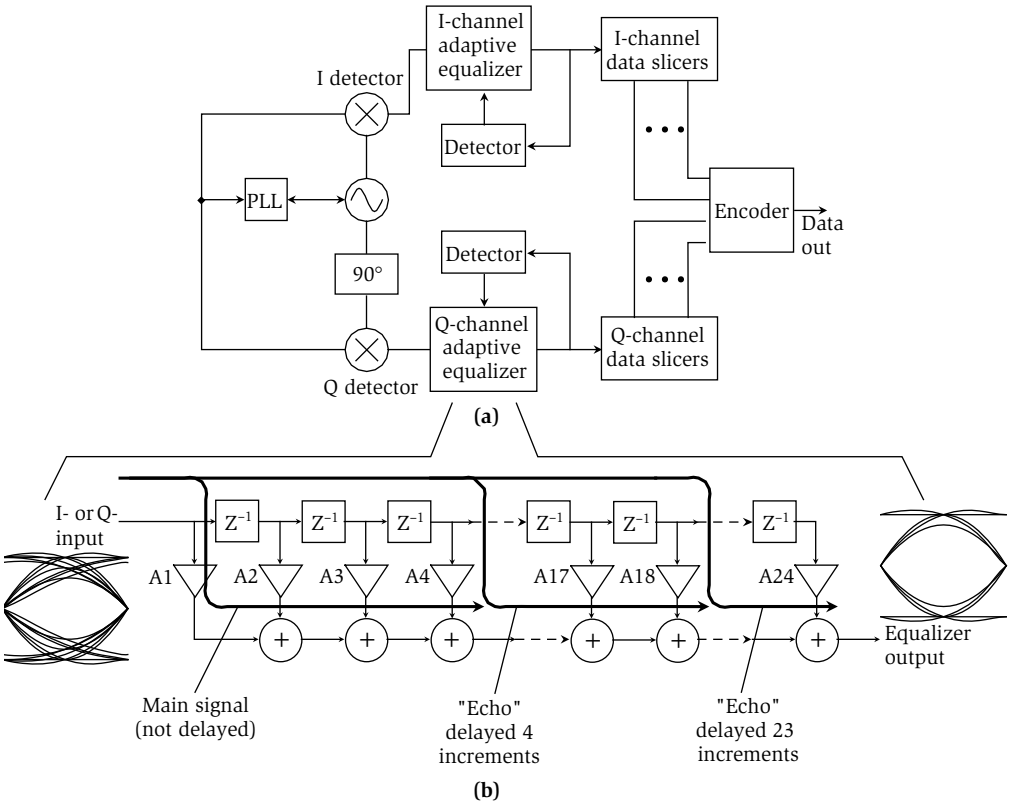


Figure 4.16 (a) Location of adaptive equalizer in receiver. (b) Adaptive equalizer.

detector, shown by each adaptive equalizer, to determine the eye closure and to adjust the taps until the eye diagram is optimized.

Adaptive equalizers are useful for compensating for echo and also for compensating for group delay and amplitude response. In some instances the main tap will not be the first one, because preechoes are possible. Amplitude errors in the channel may be modeled as a pair of echoes, one before and one after the main signal, so the detector may have to assign another tap the job of passing the highest signal (highest value of A), in order to have some preecho taps to use in canceling amplitude errors. A common combination is 7 preecho, the main tap at the eighth position, and 24 posttaps.

In the past it has been most common to place the adaptive equalizer in the receiver, as is done in digital video receivers. However, DOCSIS modems have a unique need in the upstream direction that has caused the equalizer to be placed in the upstream transmitter (at the modem) and the detector in the Cable modem

termination system (CMTS). Since signals from each modem travel a partially different path, they are subject to different distortions, depending on that path. Thus, each modem will require a different adaptive equalizer setting. A detector in the CMTS monitors the eye opening for each modem and sends signals to that modem telling it to adjust its adaptive equalizer filter until the eye opening is maximized. After that has been done, the eye opening is checked on each transmission, and changes are commanded from the CMTS to the modem when necessary.

4.2.11 Defining the Level of a Digitally Modulated Signal

Analog television signals are measured at the peak amplitude of the signal, which corresponds to the sync tips. This level is constant with modulation. The formal definition of the signal level is "the RMS value of the carrier amplitude during sync tips."

Digital signals are sometimes measured based on peak level, but it is standard in cable TV practice to measure the signal based on average level. A thermocouple power meter, which measures average power, is the basic instrument used. Measurement of the average level is more accurate, because it can be measured by supplying the signal to a thermocouple-type RF power meter. The power read on the power meter is usually expressed in decibels with respect to 1 milliwatt (dBm). Such meters are simple and quite accurate. As normally used, a spectrum analyzer comes close to measuring the average signal level (after corrections), not the peak.

The formal interpretation of "signal level" as applied to digital signals is as follows.

> The definition of the level of a digital signal shall be the power level as measured by a power meter which uses a thermocouple as a transducer. That is, the measurement shall be the average power in the signal, integrated over the actual occupied bandwidth of that signal. The signal level should be presented in decibels with respect to 1 millivolt RMS in a 75-ohm system. Thus the measurement reported is the RMS value of the sinusoidal voltage that would produce the same heating in a 75-ohm resistor as does the actual signal.[6]

This definition simply means that the reported signal level is the level of a CW signal that would produce the same heating in a resistor as does the actual signal. The signal level is converted from dBm (commonly reported by power meters) to the equivalent value in decibels with respect to 1 millivolt in a 75-ohm system (dBmV), the common unit of measure of signal strength in cable TV practice. The relationship between signal level measured in dBm and that measured in dBmV, assuming a 75-ohm system, is the same as in the analog domain:

$$\text{signal level (dBmV)} = \text{signal level (dBm)} + 48.75$$

4.2.12 Bit Error Rate Versus Carrier-to-Noise Ratio

It is appropriate to discuss the bit error rate versus the carrier-to-noise ratio of the signal. Usually in data communications, though, the independent variable is not taken to be carrier-to-noise ratio but is an equivalent measurement, expressed in terms of the energy per bit (E_b) divided by the noise per hertz of bandwidth (N_0). We explain this concept and then discuss the requirements for various modulation formats.

The Concept of E_b/N_0

For a given modulation format, as one increases the data rate, the required transmission bandwidth will increase proportionally. If a 1-Mb/s datastream requires 0.5 MHz of transmission bandwidth, then a 2-Mb/s datastream will require 1 MHz of bandwidth. The bandwidth required has doubled, doubling the noise that of necessity must be admitted by the filtering, but the data rate has also doubled.

In order to compare the performance of different modulation formats, which may be used to transmit different bit rates, we use the concept of energy per bit divided by noise power per hertz, E_b/N_0. This parameter, pronounced "E B over N zero," removes data rate and noise bandwidth from the expression. It serves the same purpose as a carrier-to-noise ratio in analog video transmission, but it is not the same as carrier-to-noise ratio. One can think of this parameter as the average carrier-to-noise ratio per bit. In a given case, one can convert to carrier-to-noise ratio by multiplying E_b/N_0 by the bit rate and then dividing by the noise bandwidth. In logarithms, this is done as follows:

$$\frac{C}{N} = \frac{E_b}{N_0} + 10\log(\text{bit rate in b/s}) - 10\log(\text{noise bandwidth in Hz})$$

The use of E_b/N_0 permits one to compare the performance of different modulation systems without having to correct for different bit rates between the systems.

BER Versus E_b/N_0

The bit error rate of various levels of QAM modulation can be approximated by the equation[7]

$$P_M = 2\left(1 - \frac{1}{\sqrt{M}}\right) \operatorname{erfc}\left(\sqrt{\frac{3k\gamma_b}{2(M-1)}}\right)\left[1 - \frac{1}{2}\left(1 - \frac{1}{\sqrt{M}}\right) \operatorname{erfc}\left(\sqrt{\frac{3k\gamma_b}{2(M-1)}}\right)\right],$$

where

M = number of states per symbol (e.g., 64 for 64QAM)
γ_b = average C/N ratio per bit (E_b/N_0)

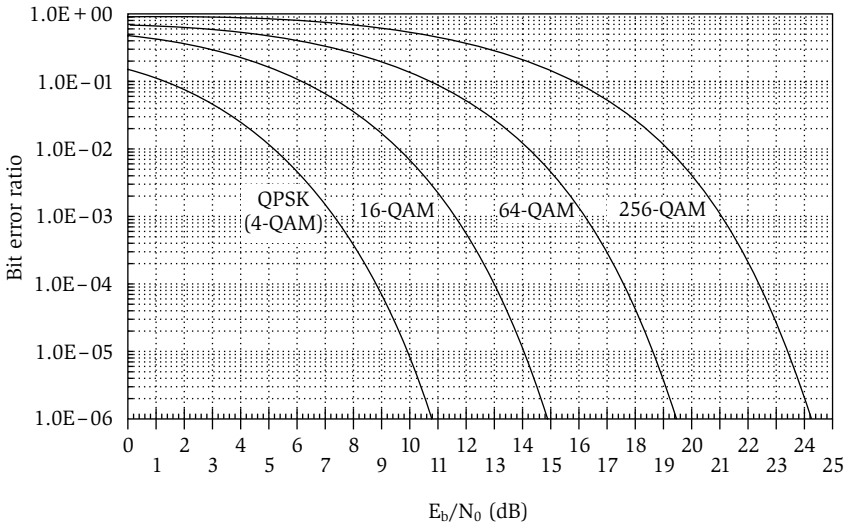


Figure 4.17 Bit Error Ratio vs. E_b/N_0 for Several Levels of QAM.

The error function, erfc, is a mathematical expression of the probability that a normal random variable (such as random noise) will exceed a certain level.

Graphing this expression for a number of levels of QAM yields the chart of Figure 4.17. This graph is normalized for a carrier-to-noise ratio, expressed as E_b/N_0, the energy per bit divided by the noise per hertz of bandwidth. This is a common expression for carrier-to-noise ratio, which is independent of bit rate and bandwidth. This graph does not include the effects of other impairments from the signal source, the transmission channel, or the receiver.

Note that digital communications engineers often refer to this as *signal*-to-noise ratio rather than *carrier*-to-noise ratio. Cable television engineers often reserve the term signal-to-noise ratio, for the baseband measurement and use carrier-to-noise ratio for the RF measurement. Digital communications engineers don't always make this distinction.

4.3 Forms of Spectrum Sharing

This section discusses several forms of spectrum sharing, in which multiple users share the same spectrum.[8] The basic reason for spectrum sharing is better efficiency of a scarce resource (spectrum). Some of the forms shown also claim advantages in the efficacy of data transmission. Note that these forms do *not* represent new modulation techniques — they generally use some level of QAM

modulation — but they do divide the spectrum differently. Nonetheless, some writers have used the term *modulation* to describe them. The careful reader will note similarities among some of the techniques. Differences may sometimes be subtle.

4.3.1 Time Division Multiple Access (TDMA)

One mature method of sharing spectrum is time division multiple access (TDMA), which is used extensively in data systems, thanks to its relative simplicity and maturity. It is quite useful for return path transmission, in which a number of return transmitters need to communicate to the headend.

Figure 4.18 illustrates time division multiple access. The reference system is a number of transmitters in one node, each of which must communicate back to the headend. In TDMA systems, each transmitter transmits on the same frequency but at different times. After each transmitter has its turn, the sequence begins again. One set of transmissions from each transmitter is called a *frame*. Ideally the level of each transmitter is adjusted such that a consistent receive level is measured at the headend. TDMA systems are covered in more detail in Chapter 6.

Some guard time must be allowed between individual transmissions, to allow for errors in the transmission time of each individual transmitter. The receiver must also be able to lock to each transmission when it starts. Of course, since each transmitter is not transmitting full time, the transmission rate must be fast enough to allow for each transmitter to transmit its complete set of bits during its turn. The bandwidth required is increased accordingly. A synchronizing signal must be sent from the headend, which defines the time slots.

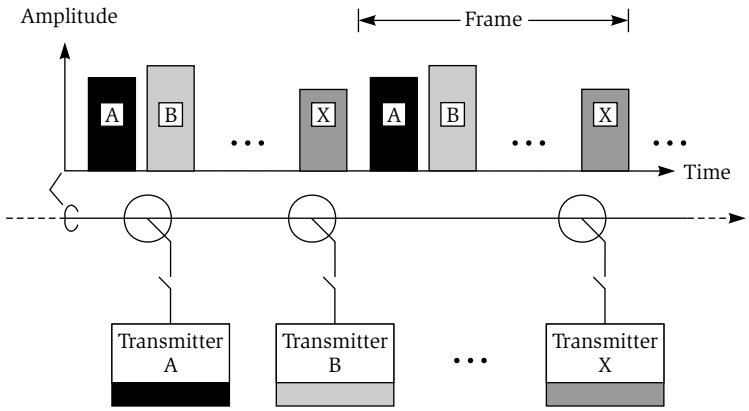


Figure 4.18 Time Division Multiple Access.

It is not necessary that each time slot (A, B, and X in Figure 4.18) be of equal length. Several protocols, including DOCSIS, allow the length to be varied, depending on the amount of data to be transmitted. This is explained in Chapter 5.

For downstream signal transmission, where only one transmitter is involved, a similar system can be used. The only difference is that it is not necessary to turn off one transmitter and turn on another. This transmission method is called *time division multiplexing*, or TDM. It is covered in more detail in Chapter 6.

4.3.2 Frequency Division Multiplexing (FDM)

Another long-established method of sharing spectrum is frequency division multiplexing (FDM), as illustrated in Figure 4.19. FDM systems use a separate narrow subchannel for each transmission, dedicating the channel to that transmitter. Compared with TDMA systems, FDM requires multiple receivers at the headend, though arguably digital signal processing (DSP) techniques significantly ease that burden. (DSP uses a specialized microprocessor to implement signal-handling functions digitally. Modulation and demodulation functions, for example, can be described mathematically. Traditional circuitry approximates the appropriate mathematical functions. DSP techniques directly implement the appropriate mathematics.)

A guard band is required between channels, to allow for practical filtering limitations. This causes some loss of efficiency, though arguably TDMA systems also exhibit some loss of efficiency due to the need for guard times between transmissions. Any suitable modulation method may be used by each transmitter. It is possible to monitor the error rate on each and to vary the modulation format to suit the quality of the channel.

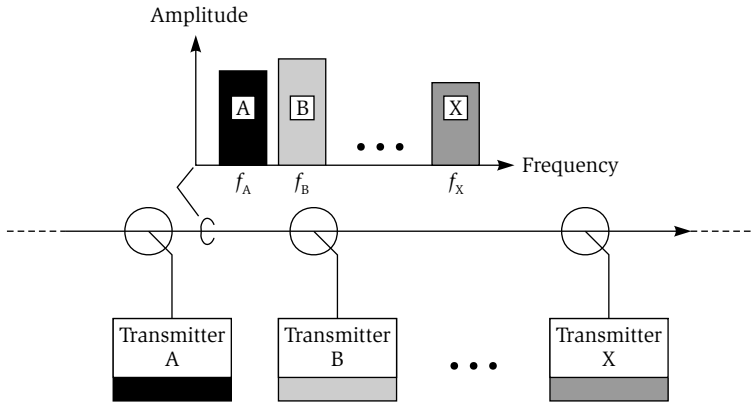


Figure 4.19 Frequency Division Multiplex.

4.3.3 Orthogonal Frequency Division Multiplexing (OFDM)

Figure 4.20 illustrates the idea of OFDM transmission. The incoming data is broken into a number of individual datastreams, each of which is then modulated onto its own carrier. The carriers are spaced just far enough apart to prevent undue adjacent channel interference. Each carrier has modulation sidebands, as illustrated by the dashed lines in the figure (separated for clarity). The carriers may be moved closer together than one would normally expect based on the data rate if the timing of bits on adjacent carriers is carefully controlled. The adjacent signals are then said to be *orthogonal*.

Any convenient modulation may be used on each of the carriers. QAM is implied in the figure. To generalize the type of modulation, communications engineers often use the term *m-ary* to refer to a modulation of any type. Note that OFDM is not itself a modulation technique, though you will often see it described as if it were.

By dividing the spectrum as shown, one replaces a single band of high-speed data with a large number of bands of lower-speed data. Since demodulation is done over a narrower band, the effects of impulse noise and group delay are reduced, and the effective carrier-to-noise ratio is improved in each channel by the narrowing of the passband.

It is possible to monitor the error rate on each carrier individually. If one channel develops an unacceptable error rate, its density of modulation can be reduced (e.g., go from 16-QAM to QPSK modulation). This will necessitate reducing the bit rate on the carrier, meaning that either some bits will have to be reassigned to another channel or the overall bit rate will have to be reduced.

For return path services, it is possible to allow bits on different subchannels to originate from different transmitters. In this case, the technique has been called orthogonal frequency division multiple access (OFDMA). It is also possible to utilize each subchannel in a time division multiple access form, where

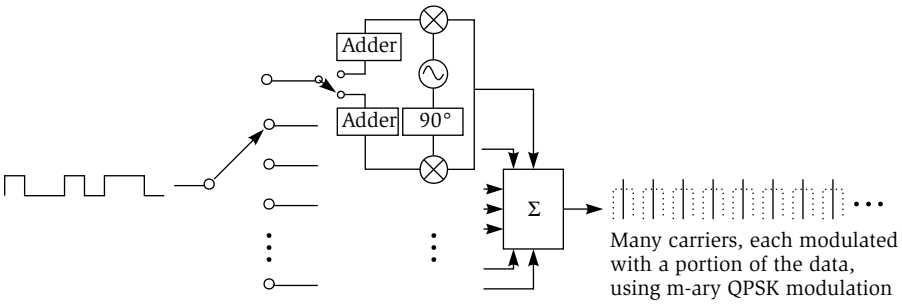


Figure 4.20 Principle of OFDM Transmission.

each subchannel is shared by many transmitters, one at a time. This has been called variable constellation multitone (VCMT).

A closely allied sharing method is synchronous discrete multitone (S-DMT). Adjacent subchannels are synchronized such that spectrum overlap is minimized, allowing closer spacing of carriers. This effectively negates the disadvantage of having to allow guard bands between channels. It does require careful synchronization between transmitters, however. Adjacent channels so synchronized are said to be *orthogonal*. Yet another similar method is known as discrete wavelet multitone (DWMT). The method again assigns small subchannels to carry data in long symbols (i.e., a low data rate in each subchannel). You might observe that some of these methods are different names for the same basic technique.

4.3.4 Code Division Multiple Access (CDMA)

Code division multiple access (CDMA) is a form of spread spectrum communications that is being used for, among other applications, some second-generation cellular phone systems (others use TDMA). The method is also known as S-CDMA, where the S stands for *synchronous*. S-CDMA is one of the advanced upstream options specified for DOCSIS 2.0.

Figure 4.21 illustrates the principle of CDMA transmission. The basic modulation is often QAM of some level. The data to be modulated on each axis is exclusive ORed with a higher-speed pseudorandom bit sequence (PRBS), which is also known as a *spreading code* or *spreading sequence*. The signal is then modulated, normally using m-ary QAM. The effect is to spread out the carrier over a wider bandwidth than necessary for carriage of the information.

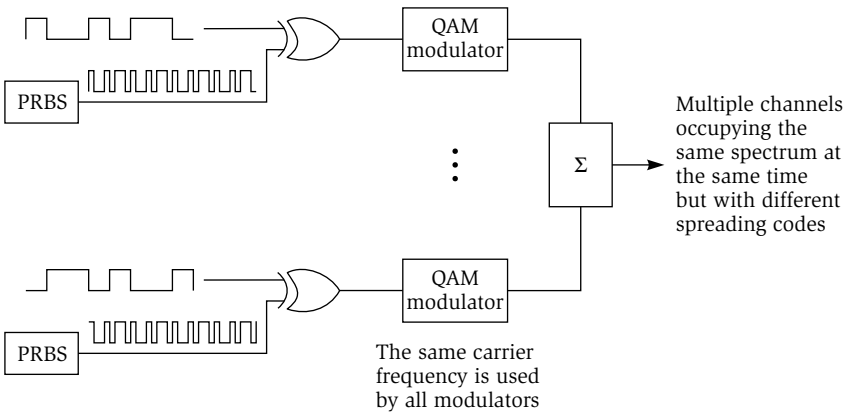


Figure 4.21 Principle of CDMA Transmission.

At the receiver, the demodulated signal is again exclusive ORed with the same PRBS. The result is a collapse of the transmission bandwidth back to that required to convey the information signal without the PRBS. Recovery of the original signal can then proceed. In the process, any narrowband interference that is added to the signal during transmission is spread, most of it out of the receive channel. In addition, if an impairment, such as a suck-out, affects part of the channel, the signal quality is minimally degraded, because most of the energy lies outside of the affected portion of the passband. (For the benefit of data communications engineers, the term *suck-out* is used in the cable television industry to mean a narrow frequency-selective reduction in the response of a transmission path.)

Figure 4.21 illustrates that many intentional transmissions may occupy the same frequency band simultaneously without interfering with one another, so long as each uses a different spreading sequence. All signals may be received simultaneously, with each having its own unique *despreading* sequence applied to separate it from the other signals. As each signal is separated with its despreading sequence, power in the other signals is spread out as if it were background noise. All spreading codes used must have the mutual (i.e., shared) property of *orthogonality*. That is, they must be noninterfering with each other. The practical limitation on the number of signals that can be transmitted in the same bandwidth is the number of orthogonal spreading codes available. In turn, the number of spreading codes depends on the number of bits in the spreading sequence: The longer the sequence, the more orthogonal codes exist. Spreading code design is a complex subject.

The type of spread spectrum system illustrated here is called *direct sequence* spread spectrum. There exists another practical spread spectrum technique, known as *frequency hopping* spread spectrum. In this technique, the frequency on which the digital signal is transmitted is moved multiple times during the transmission of one bit. The receiver follows the transmitter's spectrum hopping. One can see intuitively that a narrow band of interference present at one or a few of the frequencies in the hopping sequence will have little effect on the received signal.

With any spread spectrum approach, the total spectrum occupied by the signal is much wider than what would be occupied by the signal without spreading. Spectrum efficiency must be improved by transmitting many signals in the same spectrum at the same time.

4.4 Measuring Digitally Modulated Signals

Of great interest to cable engineers is the need to measure digitally modulated signals. Some measurements may be made using the same instruments as used in analog-modulated signals, such as a spectrum analyzer. Other measurements require new instrumentation.

4.4.1 Measuring Signal Level

Unlike an analog signal, a digitally modulated signal has no distinct peak in signal level at carrier frequencies. Most of the time, the carrier is not transmitted, at least not at full amplitude. Because digital signals are almost always randomized before being modulated, the spectrum is flat over its passband. As stated earlier, the average signal level is measured, not the peak level as is the accepted practice in measuring analog video signals. A spectrum analyzer is often used for this measurement, but several corrections in its reading must be made. The biggest is to correct for the occupied bandwidth of the signal, versus the bandwidth over which the measurement is made. The bandwidth over which the signal level is measured (the *noise bandwidth*) is roughly equal to the resolution bandwidth of the spectrum analyzer. However, the two are not quite the same, and for accurate measurements a correction factor must be used.

Since signal energy is distributed evenly across the band, the power of the signal is proportional to the bandwidth over which the measurement is made. This can be corrected to the actual bandwidth of the signal as follows:

$$S = 10 \log \frac{\mathrm{BW}_{\mathrm{noise}}}{\mathrm{BW}_{\mathrm{signal}}}$$

where

$$\mathrm{BW}_{\mathrm{noise}} = \text{Noise bandwidth of analyzer IF filter}$$
$$\mathrm{BW}_{\mathrm{signal}} = \text{3-dB bandwidth of signal}$$

This corrects for the occupied bandwidth versus the measuring bandwidth, assuming that the spectrum is flat. Additional corrections that must be made include one for how the detector in the instrument responds to noise and one for how averaging is done. The instrument's manufacturer must specify these corrections. In most modern measurements the analyzer has a noise marker measurement facility that makes these corrections for you.

The noise bandwidth we refer to in the equations is defined as the bandwidth that would be occupied by an ideally flat filter with infinitely sharp edges (*skirts*), which admits the same amount of flat noise as does the real filter.

Note that the same measurement technique may be used with QAM, OFDM, or 8-VSB signals. With the 8-VSB signal, a small carrier component is transmitted 310 kHz above the band edge, but it is so small that it makes little difference in the signal level. Instruments other than a spectrum analyzer may be used to measure the amplitude of a digital signal, but the manufacturer must provide conversion factors to correct the measured reading to the actual average signal level.

4.4.2 Measuring Bit Error Rate (BER) and Modulation Error Ratio (MER)

The quality of a digital signal may be evaluated using *bit error rate* (BER) measurements if the signal quality is low enough to introduce bit errors. If the signal quality is such that errors are not being made, *modulation error ratio* (MER) may be used as an indicator of signal quality.

Bit Error Rate

BER is simply the average number of bits received in error divided by the total number of bits received. Scientific notation is used to express BER, since the number of errors had better be quite small. If one bit out of every million received is in error, then the BER is

$$\text{BER} = \frac{1 \text{ errored bit}}{10^6 \text{ bits}} = 10^{-6}$$

Note that BER is dimensionless. In order to measure BER in a laboratory setting, usually a known bit pattern is transmitted repeatedly. The receiver compares the received bit pattern against what it knows was transmitted and then displays the BER. It can take a long time to measure BER when the BER is low. Suppose it is desired to measure the BER of a 64-QAM signal being transmitted at a rate of 30.342 Mb/s. If the BER is 10^{-10}, then we will wait, on the average, through 10^{10} bits before we receive a bad one. This will require

$$\frac{10^{10} \text{ bits}}{30.342 \times 10^6 \text{ bits/sec}} = 329.6 \text{ sec} = 5.49 \text{ min}$$

before one error is received. We must receive a number of errors before we can conclude that we have a statistically valid sample. You can see that it will take a long time to make this measurement.

In-service BER cannot be performed on a known bit pattern because, by definition, the receiver cannot know what the transmitter is sending, unless the transmitter and receiver are located at the same place and the receiver is also getting the bit stream being sent to the transmitter. There are ways of estimating BER from the received signal without having to know the transmitted pattern. One way is to look at the eye diagram (see Figure 4.11). As more interference, noise, or other impairments enter the transmission path, the open center of the eye closes. The BER may be estimated from the amount of eye closure.

Another way to estimate BER is to look at the number of times errors are corrected. Most digital systems transmit extra bits for the purpose of correcting errors in the transmission (see Section 3.6.2). When these corrections are made, an error is counted. The bits received are also counted, and the ratio becomes the BER.

BER may be measured before or after error correction. These measurements are sometimes called the *raw*, or *precorrected, BER* (measured before error correction) and the *corrected*, or *postcorrection, BER* (after error correction). Both measurements have their place. Precorrected BER is the better indication of the channel performance, but postcorrection BER is the better indication of the quality of the signal the user will experience.

Modulation Error Ratio

Figure 4.22(a) plots bit error rate against transmission quality. The transmission quality includes degradation due to noise, distortion, frequency response errors, group delay, echo, gain compression, and other things that can damage the quality of the digital signal. In the left region, where transmission quality is low, the BER specifies the quality of the signal. However, above a certain noise density, the BER is so high we either don't see errors or at least don't have the patience to wait long enough to measure them. Even in this region, however, we are interested in knowing the quality of the channel. Since the change in transmission quality between a perfect picture and no picture is very small, it is possible that a technician will leave a home with perfect picture quality, only to have the subscriber call a few minutes later with no picture at all. Thus, it is not enough to know that a picture is good or that data is being transmitted efficiently. We need to know the quality of the channel even when things are going well.

A useful metric for communications quality is the *modulation error ratio* (MER). MER is defined in Figure 4.22(b), which shows one quadrant of the constellation diagram for a 16-QAM signal. Specialized test equipment is used to display the constellation diagram and to measure the MER. In Figure 4.22(b) the nominal location of each point in the constellation is shown as an *open circle*. Several pluses mark locations that the constellation point may occupy

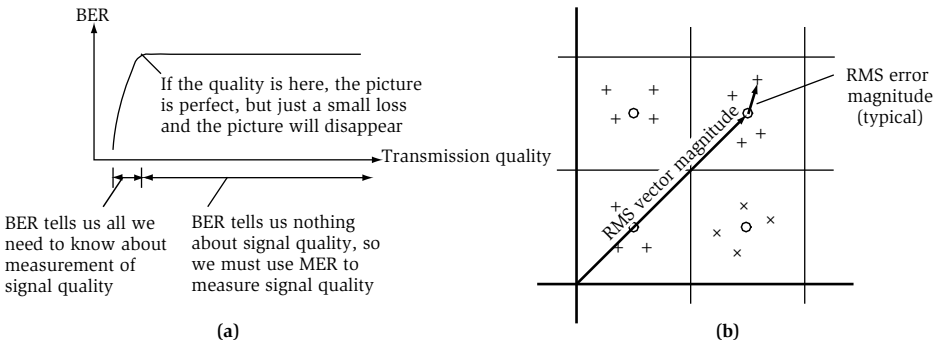

BER
If the quality is here, the picture is perfect, but just a small loss and the picture will disappear
Transmission quality
BER tells us all we need to know about measurement of signal quality
BER tells us nothing about signal quality, so we must use MER to measure signal quality
(a)
RMS error magnitude (typical)
RMS vector magnitude
(b)

Figure 4.22 (a) BER vs. transmission quality. (b) Definition of MER.

for some one particular transmitted symbol. In practice, the whole area around the nominal location of the point will be occupied with points, because channel impairments affect each symbol differently. Some common channel impairments are shown in Figures 4.12 and 4.13. MER is a single number used to evaluate the channel quality regardless of the source of impairment.

MER is defined as 10 times the log of the ratio of the RMS error magnitude to the RMS vector magnitude to the nominal position of the point:

$$\text{MER} = 10 \log \frac{\text{RMS error magnitude}}{\text{RMS vector magnitude}}$$

The error magnitude and the vector magnitude are defined in Figure 4.22(b). Note that, while the RMS vector magnitude is constant for each point in the constellation, the RMS error magnitude will change for each symbol, since random processes are at work. In order to derive the MER, the test instrument sums the RMS error magnitude over many samples, at each point in the constellation, and averages them.

MER measures the "fuzziness" of the constellation. MER is sometimes described as similar to C/N measurements in analog transmission. While there are similarities (both are measured in decibels and higher numbers are better), there are also differences. Many things other than just poor C/N contribute to degradation of MER. Any transmission channel impairment affects MER, including random noise, group delay, echoes, phase noise, and gain compression.

In order to achieve good performance with 64-QAM, a MER of at least 23 dB is required. For 256-QAM, a MER of at least 28 dB is necessary. A minimum safety margin of 3 dB above these numbers is recommended. Portable instruments available tend to have a maximum BER measurement capability of about 34–38 dB.[9]

A related measurement you may see sometimes is the *error vector magnitude*, defined as the ratio of RMS error magnitude to the RMS vector magnitude, expressed as a percentage:

$$\text{EVM} = 100 \frac{\text{RMS error magnitude}}{\text{RMS vector magnitude}} \%$$

The impairments mentioned earlier have different effects on MER and BER. Noise (see Figure 4.17) is the most studied and is well documented. Phase noise and compression affect mostly the outer points of the constellation, so the MER is not affected much, yet BER may be degraded seriously. CW interference affects both MER and BER because all constellation points are impaired. Impulse noise creates havoc with the BER, but because of its transient nature the average impairment is small and MER is still good. Composite second order (CSO) and composite triple beat (CTB), because of their peaks, look more like impulse noise than CW interference. Each impairment has its own "power," and the MER reflects the sum of these powers.[10]

Because digital transmission is not perfect, transmission errors do occur. To render the system operational, *channel coding* is used to mitigate such transmission errors. A good example is the downstream digital video or cable modem transmission standard, ITU-T J-83 Annex B, which shows three levels of channel coding.

The best-known channel coding is *forward error correction* (FEC). The data words are grouped in packets, to which are appended Reed–Solomon (R-S) error-correcting words, e.g., 122 data words for six R-S words (7-bit words). At the receiver the R-S codes are used to detect almost all errors and, in addition, can correct one error for each two R-S codes (in this case up to three errors per packet). This is the basis of the pre-FEC and post-FEC bit error rates.

The second channel-coding mechanism is *interleaving*. Instead of transmitting all the words of each packet in sequence, they are interleaved. That is, adjacent transmitted words comes from different packets. For instance, all the first words of 128 packets, then all the second words, and so on until all the 128th words of all 128 packets are transmitted. At the receiver, the transmitted words are *deinterleaved*, to be regrouped in their original packet order. If impulse noise causes errors on many consecutive transmitted words, then without interleaving the R-S could not correct all errors. With interleaving, the grouped errors in transmission are now spread over a larger number of packets, each having only a few correctable errors. Interleaving provides a *burst protection period* proportional to the length and number of packets interleaved. On the other hand, it introduces a delay, or *latency*, in transmission and reception due to the storage requirement of the packets. This is why cable modem specifications limit interleave to 128×1. Digital video is more tolerant of latency, and deep interleave of 128×4 is often used (that is, the interleave is done over 128×4 packets, each of 128 words). The protection period is longer, at the expense of larger latency.

The third channel coding present in Annex B is *trellis code modulation* (TCM). TCM does not allow all combinations of symbol sequences. Rather, it limits the choice of the following symbol to specific ones according to the preceding symbols and specific rules. At reception, the same rules are used to select *the most probable word sequence*. There is no guarantee that it's the right one, but it is likely to be. Many errors are corrected, but some are not, activating the second line of defense, the Reed–Solomon FEC. Because it gives the most probable sequence but no decisive error indication, we cannot measure a pre-TCM bit error rate. Finally, Annex A and Annex C do not have TCM and consequently have a worst pre-FEC error rate than Annex B.[11]

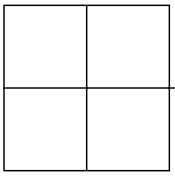
4.5 Summary

This chapter has presented the theory of digital modulation as used in cable TV and broadcast systems. The most commonly used family of modulation formats are quadrature amplitude modulation, QAM. The number preceding the QAM designation indicates the number of states the digitally modulated signal can

assume, and ranges from 4-QAM (also known as QPSK) to 256-QAM. There may be future applications for even higher levels of QAM. The next chapter will address a companion topic, protocols used in DOCSIS and other digital transmission standards of interest to cable operators.

Endnotes

1. Advanced Television Systems Committee, A/54, *Guide to the Use of the ATSC Digital Television Standard*, October 1995. Available at http://atsc.org/ (note: no "www").
2. Institute of Electrical and Electronics Engineers, *The Authoritative Dictionary of IEEE Standards Terms*, IEEE 100, 7th ed., 2000.
3. Society of Cable Telecommunications Engineers, ANSI/SCTE 22-1 2002, *Data-Over-Cable Service Interface Specification, DOCSIS 1.0 Radio Frequency Interface (RFI)*, Table 4-12.
4. R. D. Katznelson, *Delivering on the 256-QAM promises*. Cable-Tek Expo 2002.
5. Yasuhiro Ito, Performance of OFDM Transmission System — Peak-to-Average Power Ratio and Clipping Effect. *SPECS Technology*, June/July 1995, Cable Television Laboratories.
6. National Cable Television Association, *NCTA Recommended Practices: Upstream Transport Issues*. Though the citation is oriented to return path issues, the definition is generally applicable.
7. John G. Proakis, *Digital Communications*. New York: Mc Graw-Hill, 1983, p. 187.
8. Portions of this information are supplied by M. Bugajski of Arris Interactive, Duluth, GA., in a summary prepared for internal use.
9. Sunrise Telecom, *Modulation Error Ratio Demystified*, available at www.sunrisetelecom.com.
10. Sunrise Telecom Broadband On-line Learning, *MER, BER and More*, available at www.sunrisetelecom.com.
11. ITU-T J-83, Digital Multi-Programme Systems for the Television and Data Services for Cable Distribution.



Chapter 5
Cable Networking Protocols

5.1 Introduction to Protocols

This chapter introduces you to the concept of data protocols, with an emphasis on those used in DOCSIS modems. A number of other protocols are covered because they find frequent use in cable TV work, but they were not developed specifically for that. You will not become a protocol expert by reading this chapter, but you should be able to understand the basic concepts.

The word *protocol* is generally not used in analog communications, but it could be applied. When two pieces of equipment communicate with each other, there must be an established set of rules for that communication. In digital communications, the set of rules is called a *protocol*. For example, if we were to apply the word to the analog world, we might say that the protocol for FM radio broadcasting in monaural would be as follows:

> Deviation: 75 kHz peak
>
> Preemphasis: 75 μs
>
> Frequency: channels spaced 200 kHz apart, beginning at 88.1 MHz, through 107.9 MHz (North America)

This would define reasonably well a protocol to which monaural FM receivers could be designed, with assurance that a receiver built by one company would be compatible with a transmitter built by someone else. Data communications protocols are more complex but are intended for the same purpose: to make sure that a talker will be understood by the listener.

5.1.1 The Concept of Protocols

Data protocols can be introduced using the analogy of a letter. A letter is written by one person and intended to be read by another. In the process of sending the letter from the writer to the reader, it is handled by a number of different people and packaged in different ways with other letters. Those carrying the letter have no interest in, and no need to know, the content of the letter, which is "hidden"

in an envelope. Similarly, a digital message from one entity to another is "packaged" with other messages and transported by a transport system that has no knowledge of the contents of the message.

As shown in Figure 5.1, the sender of a letter addresses it to the intended recipient. The letter is placed in a bag with other letters for transport to a sorting center and then rebagged with other letters going to the same general destination. The bag is placed on a conveyance (truck, train, or airplane), which carries it close to its destination, where the process is reversed, before the letter is ultimately delivered to the intended recipient, who removes it from the envelope and reads it. In data transmission, the "envelope" is replaced, typically, with a *header* and sometimes a *postheader*. Subsequent handling steps pay no attention to what is enclosed, nor would they know what to do with the contents if they did look.

Just as a letter can be placed in an envelope, the envelope placed in a bag, and the bag placed in a train, different layers of the protocol may place their own header and postheader on the message. Those headers are interpreted by the corresponding process at the receiving point and discarded. The message is then passed to the next process.

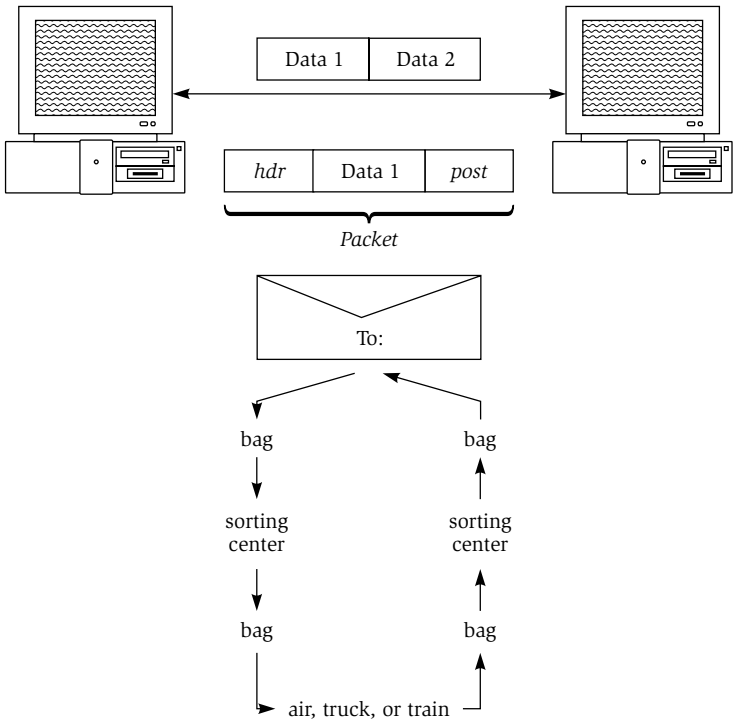


Figure 5.1 Concept of a data protocol.

A communications protocol defines the set of steps a message takes from the sender to the receiver. Analogous to the letter concept of Figure 5.1, the protocol defines how a data signal is handled at each stage of transmission. One layer knows nothing about how the message is transmitted at a different layer. Just as a letter-sorting process has no interest in whether the letter is to be transported by airplane or truck, one layer of a protocol has no interest in how the message is to be handled at another level.

One advantage of using the layering concept is that large portions of the system can be designed such that data can be carried on a number of different media. For instance, a data message can be constructed the same whether it is to be sent over a short point-to-point wired link or over a satellite link shared with a lot of other users. This enables efficiencies in writing software and constructing hardware.

5.1.2 The OSI Seven-Layer Reference Model

A common way to explain the role of different protocol levels is the so-called seven layer model promulgated by the International Telecommunications Union (ITU). The model is known as the *O*pen *S*ystems *I*nterconnection (OSI) model.[1–3] 'ISO' is sometimes used, for the International Standards Organization. Not all modern protocols fit neatly into the seven layers. Nonetheless, the model provides a useful frame of reference for understanding protocols.

Figure 5.2 illustrates the OSI seven-layer model and includes examples of common protocols and approximately where they would fit into the model. The model involves transmitting information from, for example, the left column (sender) to the right column (recipient). At the top of each column ("stack") is the application, such as a Web browser or an email system. The bottom three or four layers are usually of the greatest interest to engineers working on data transmission. These bottom layers define the transmission network, whereas the higher layers define the end user system.

You can think of Figure 5.2 as describing the flow of information in a computer or other device, from the application (Web browsing, electronic program guide, etc.) through to the communications port on the device. The information is transported to whatever machine is talking to the first and flows up again. In the middle is the communications network, which involves up to layer 3 or maybe 4. Note that different protocols may exist in the end applications and the communications system. The same protocol is not necessarily used in the end machines as in the communications network. This is fine as long as the devices can interface using a common protocol. Later we describe some of the more common transport protocols you will encounter.

Physical Layer

The lowest layer is the *physical* layer, which defines the functions required in order to get data from one place to another. Besides specifying such things as the

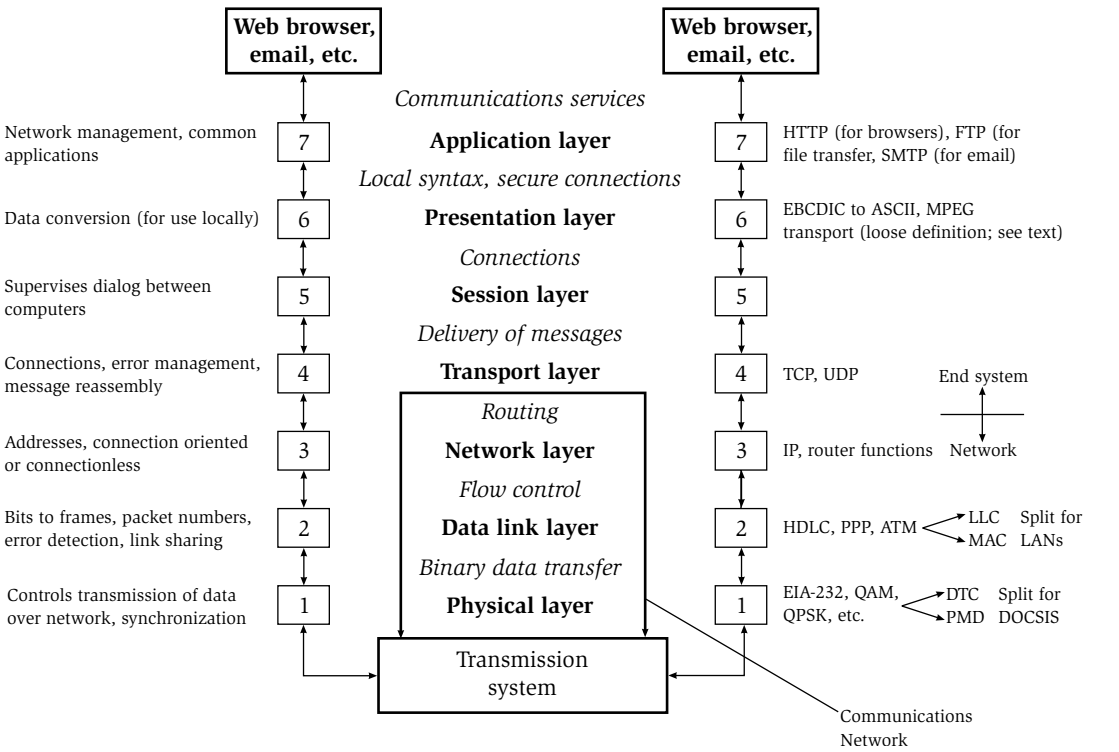


Figure 5.2 The OSI seven-layer model.

modulation format (to which Chapter 4 was dedicated), the physical layer takes care of such functions as synchronization. *Synchronization* refers to the process of ensuring that data recovery at the receiver is timed properly with the data being transmitted. The actual transmission system, such as an HFC network, is not considered part of any layer of the protocol, but obviously the protocol must have knowledge of the function below (and above) itself.

Data Link Layer

The *data link layer* handles such tasks as gathering up sets of bits for transmission as packets and making sure the packets get from one end to the other. In addition, recognizing that physical layer transmission sometimes introduces errors, the data link layer handles error detection (and sometimes correction).

When the OSI model was first introduced, data communication was point to point. With time, data communication has become much more multipoint to multipoint, so the data link layer was divided to recognize multipoint needs. With the split of the data link layer, the traditional functions of the layer went to the upper portion, called the *logical link layer*. A new sublayer, media access control (MAC), was defined. The MAC layer routes packets from a sender to a receiver along a common path. It makes sure the message arrives at the intended recipient. The MAC layer adds a physical address, defining the intended recipient machine, and controls shared access to a resource. For example, control of TDMA operation (see Section 4.3.1) is a MAC layer function.

Network Layer

The network layer handles routing of packets over a network that may consist of many links. It figures out, for each packet, the best of the available routes over which to send a packet. This is explained in more detail in Section 5.1.4.

Transport Layer

The transport layer traditionally is an end user layer and not a part of the communications network. However, as we shall see, modern communications networks are often cognizant of the layer 4 protocol, at least in order to allow for quality-of-service handling of packets.

End System Layers

The upper layers of the protocol are important, but they are treated lightly here because the emphasis is on the lower layers of the protocol, with which communications engineers tend to interact more frequently. The transport layer combines multiple low-data-rate channels into a higher-data-rate channel and then breaks down the channels at the receive end. It defines the destination for each channel. In a connection-oriented service, the transport layer sets up the connection, delivers error-free messages in the correct order, and ends the connection when it is no longer needed. In a connectionless service, the transport layer delivers messages one by one, possibly with errors and with no guarantee of the correct order.

The session layer supervises the dialog between two computers. It is used, for example, in the transmission of audio and video channels in a video conferencing session. Synchronization points are inserted between large numbers of messages, such as the continuous messages of video and audio files. Synchronization points are packets that divide the sequence of messages into groups.

The presentation layer converts between different representations of data. For example, some computers represent words by placing the most significant

bit first, whereas other computers do the opposite. The presentation layer adapts the data to be used on the host machine if necessary. MPEG video encoding may be loosely thought of as a presentation layer function. The MPEG transport layer probably would be better thought of as a combination of the transport layer and some functions from lower layers. (MPEG does not really conform to the seven-layer model.)

The application layer interfaces with end application programs that require communications functions. The application programs, such as Web browsers, don't *reside* in the applications layer themselves; they are above all the communications services. The application layer provides file transfer, terminal emulation, remote log-in, directory service, and remote job execution. These services are used by the end application.

5.1.3 MAC Layer Issues of Interest in Cable Modem Applications

Of particular interest to users of cable modems is the MAC sublayer of the data link layer. It falls to the MAC layer to provide an efficient protocol to allow operation of modems in the HFC environment. This physical environment is different from that of many traditional data networks. A significant issue is that each upstream "talker" cannot "hear" other upstream talkers unless the upstream transmission is repeated in the downstream direction. Also, unlike local area networks, the propagation delay on the HFC plant is significant and, if not taken into account, can create inefficiencies in data communications. The following discussion of the use of time slots presupposes a time division multiple access (TDMA) protocol in the upstream direction. It serves as an introduction to the more detailed section 5.2 on cable modems, which follows.

Reservation of Time Slots

In TDMA systems, some method must exist to allow for determination of which of many transmitters gets to transmit in a given time slot. One such method is generally known as *reservation*. In a reservation protocol, each talker who wants to talk requests, and is then assigned, one of the available time slots. When that time slot arrives, he or she talks and then is quiet while others talk. The management and assignment of the time slots comes from some central authority, which in the case of a cable modem system would be the headend *cable modem termination system*, CMTS. A reserved time slot allows a modem to transmit a known amount of information in a known time period. However, if a modem does not wish to use its time slot, the time slot is not used by anyone else until the controller decides that the time slot is available and assigns it to someone else. This can result in inefficiencies when there are a lot of modems, each with little to say. On the other hand, when a modem has a lot to say, reservation can be more efficient than the alternatives.

Contention for Time Slots

An alternative is to have the modems contend for time slots: When a modem needs a time slot, it simply speaks up and requests the time slot. In the case of cable modems, where one modem cannot "hear" any other modems, the talking modem will have to wait until the headend can acknowledge that it received the transmission. If another modem tried to talk at the same time — even if only part of the transmissions overlapped — then neither would be heard, so neither would be acknowledged. The transmissions are said to have *collided*. When that happens, both modems wait for an arbitrary time and then try to transmit again. If they again collide or if one collides with any other modems, they must again back off for a random time and try yet again. With some protocols, such as Ethernet, the talkers can hear each other, so they know when a collision occurs.

When talkers have relatively little to say but need to say it quickly when they do have something, contention works fairly well. In cable systems, though, contention is particularly difficult, because one talker cannot hear other talkers unless the upstream transmission is repeated in the downstream direction. This was done in some early (pre-DOCSIS) cable modem systems, which were adaptations of Ethernet. However, the propagation delay of a round-trip on cable plant can be long compared with the length of a message. Consequently, by the time two modems know that they have collided, a lot of time has been wasted. In addition, the need for upstream data to be repeated in the downstream direction is an inefficient use of the spectrum resource.

Combination Protocols

Contention protocols are sometimes called *aloha* protocols. Simple aloha protocols do not feature time slots, but they do allow any modems that need to talk to do so at any time. If the protocol follows the discipline of time slots into which any talker's transmission falls, it is called a *slotted aloha* protocol. If some time slots are available for reservation while others are available for contention, the protocol might be called a *reserved aloha* protocol.

The difficulties with contention systems on cable might lead one to prefer some sort of reservation system. It is certainly possible to reserve for each modem on a particular channel a permanent time slot for upstream transmission. However, such a permanent assignment is a problem when each modem is in use for at most a few hours each day and has little to say when it is used. The rest of the time, the time slot cannot be occupied by other users. Some sort of method should be used to assign time slots on an as-needed basis so that if a modem was not using its time slot, that slot could be used by another modem.

If a modem is assigned a time slot only when it needs one, then there must be some way provided to allow a modem to request a time slot. Since the headend can have no a priori knowledge of when a particular modem will need a time slot, it is necessary to allow a modem to request one. This request implies

that there should be some mechanism of contention in a cable modem protocol. The mechanism can be one where a time slot is allocated to all modems not presently talking, in which they can request a time slot. They must do this in a contention mode. A modem needing a time slot waits until the contention time slot, and then makes a request to the headend. If it is the only modem doing so, its request will be acknowledged in the return datastream, and a time slot will be assigned if available. If a second modem has also transmitted a request at the same time, then neither will be acknowledged, and they must both back off, waiting for a random time before trying again in the contention time slot.

5.1.4 Connection-Oriented and Connectionless Service

A communication link may be differentiated on the basis of whether it is a connection-oriented service or a connectionless service. Connection-oriented services are better known to the public, as represented by the public switched telephone network (PSTN). Connectionless service is a less familiar concept. Both are described here.

Connection-Oriented Service

Figure 5.3 represents a connection-oriented service, such as telephone service. It shows a section of a switch, which couples one user (users A through H are shown) with any other user. A switch may be thought of as a matrix, with users connected to horizontal lines and vertical lines used to interconnect users as

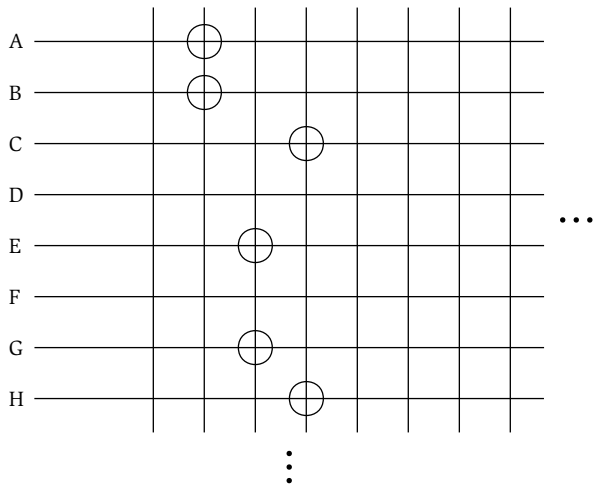


Figure 5.3 Representation of a connection-oriented service.

demanded. The circles represent connections between the horizontal and vertical lines. As shown, A is connected to B, C to H, and E to G. D and F are not in communication with any users.

A connection must be set up whenever communication is desired between any two points and then "torn down" when the communication is completed. Because time is required to set up (and tear down) a communications channel, the user must wait until a channel is set up before he can use it. Resources are dedicated to the communications channel whether or not communication is taking place at the moment. For example, during a voice telephone call, a channel is dedicated to that call even when one party has left the phone to get some information. Resources are tied up for both directions of the call even though each direction is used no more than half the time.

Connectionless Service

By contrast, a connectionless system passes information when information is available; the rest of the time, the resources are available for other users. Each packet, or "chunk", of information is treated on its own, without relation to anything else. There is no setup or teardown of a connection. Figure 5.4 illustrates the concept of routing, a key component of a connectionless circuit. The

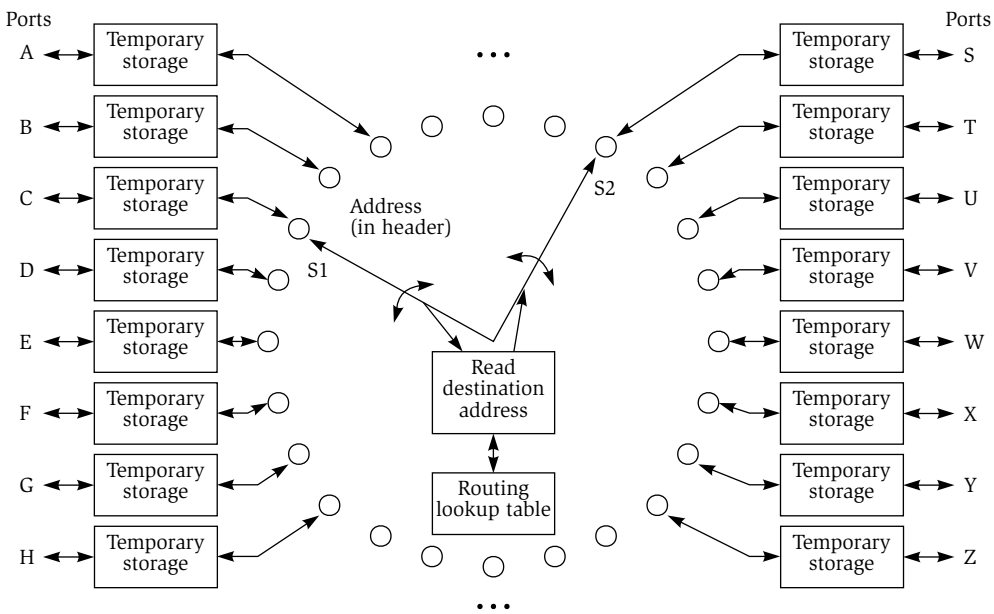


Figure 5.4 Router, representing a connectionless system.

concept is used in the Internet backbone and other IP systems. The figure shows communication ports A through H and S through Z, each connected to a temporary storage facility, which accepts a message from the user and stores it for a short time until it can be routed to the next destination. Each interface can be thought of as either a user or another point in a multipoint route between two endpoints (users). All communications routes are two way.

A rotary switch has two commutators, S1 and S2, each of which can independently connect to any interface. Conceptually, S1 connects in turn to each port. If there is data there to be routed further, that data will contain a header that defines the destination. The destination header is read, and a routing lookup table is consulted. The lookup table tells the router which of the ports moves the signal most expeditiously toward its destination. Commutator S2 connects to this port, moving the signal toward the next point on the journey to its destination. (Real routers don't use switches to make connections, but this is a convenient mechanism with which to illustrate the function.)

If A is talking to B (as in Figure 5.4), then at some time S1 will connect to A, the destination address B will be read, and the routing lookup table will be consulted. In this case, the lookup table would say to connect to B directly, since the destination port is connected to the router.

Figure 5.5 shows a larger view of a network. Each of the circled letters represents a router, as in Figure 5.4, each in a different location. Each router is connected to a number of adjacent routers. The routing table in each must know of every ultimate destination and must know how to efficiently route a packet toward that destination (this concept is modified in the case of some modern protocols, such as MPLS). In several cases, multiple routes are possible, so if one

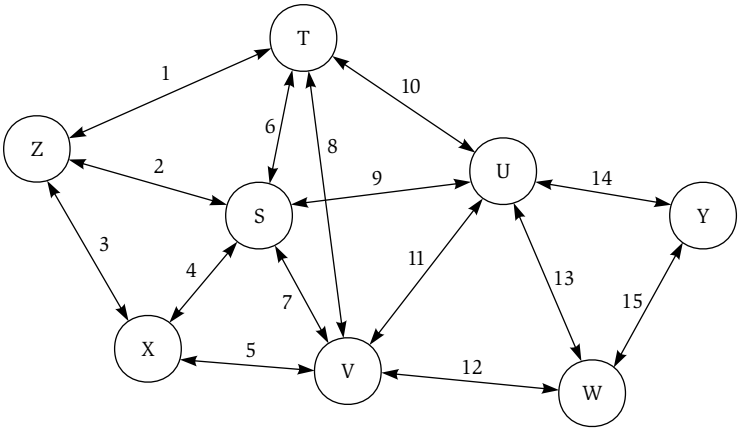


Figure 5.5 Multiple point routing.

is not available another can be used. For example, to send a packet from Z to X, the most efficient route is 3, which connects the two directly. If route 3 is not available, then the packet could be sent to S via route 2 and then to X via route 4. Route 3 might not be available because it has failed, been taken out of service, or lacks the capacity to handle all of the traffic being fed to it at the moment. The routing table in each router must be aware of the best route and alternates, and must be aware of the present condition of each route.[4]

Differences Between Connection-Oriented and Connectionless Service

In summary, connection-oriented services, an example being the public switched telephone network, dedicate a communications path to two users for the duration of a communications session. The session may last less than a second, or it may last for days. When communication is desired, a circuit must be set up that is dedicated to that communicating pair for the duration. At the end of the session, the circuit is torn down, releasing any common resources for other users. The device at the heart of a connection–oriented service is called a *circuit switch*.

In contrast, connectionless circuits are established on an ad hoc basis when they are needed for an individual packet of information. As soon as the packet passes, the resources are available for other users. Each packet of information has a destination address. That destination address is read by the router, which then consults a routing table to learn the best route on which to send the packet at that moment.

To the user of a connection-oriented service, it is necessary to establish the circuit (e.g., by placing a phone call), wait for the connection to be established, and then use it. The delay, once the circuit is established, is short and predictable. A user of a connectionless service is not aware of the establishment of a path for each packet, that function being handled automatically in protocol. So far as the user is concerned, the communication path is always available to him. On the other hand, because of route diversity, sequential packets may travel on different paths; and since some routes may be faster than others, it is not guaranteed that packets will arrive on a regular basis or even in the right order. It is the responsibility of the protocol to rearrange packets into the correct order and to present them to the using function correctly. (With reference to the OSI model of Figure 5.2, this function would be handled by the transport layer.)

Connectionless services use system resources more efficiently because no resources are tied up between users who have established a communications path they are not actively using at the moment. On the other hand, the user of a connectionless service “thinks” a communications channel is available at all times, though he or she may experience delays. Of course, many of the delays are small in human terms, and packets received out of order may be properly ordered without user awareness.

5.2 The DOCSIS Protocol for Cable Modems

Early cable modem systems used proprietary protocols, many derived from Ethernet. A group of cable television operators and Cable Television Laboratories (CableLabs) collaborated in the development of a set of purchase specifications to be used by participants and other interested parties as a procurement standard for cable modems. The collaboration initially operated under the name MCNS Holdings, L.P., an acronym for Multimedia Cable Network System. The set of specifications is referred to as DOCSIS, for the formal name of the specification set, data-over-cable service interface specification.[5] The DOCSIS standards were later adopted by the Society of Cable Telecommunications Engineers (SCTE), a recognized standards body. Subsequently they were adopted as national standards by the American National Standards Institute (ANSI).

DOCSIS was the first opportunity to allow cable TV equipment to be owned by the consumer. Standardization is necessary to ensure that any modem bought by any consumer can connect to the network. Further, the subscriber can assume that if he or she moves to another cable system, the same modem will work in the new system. The standard allows multiple vendors to make competing products. With competition comes innovation and lower costs.

5.2.1 DOCSIS Reference Model

Figure 5.6 illustrates the headend-to-home portion of the DOCSIS reference architecture.[6] The architecture is fairly straightforward. In the downstream

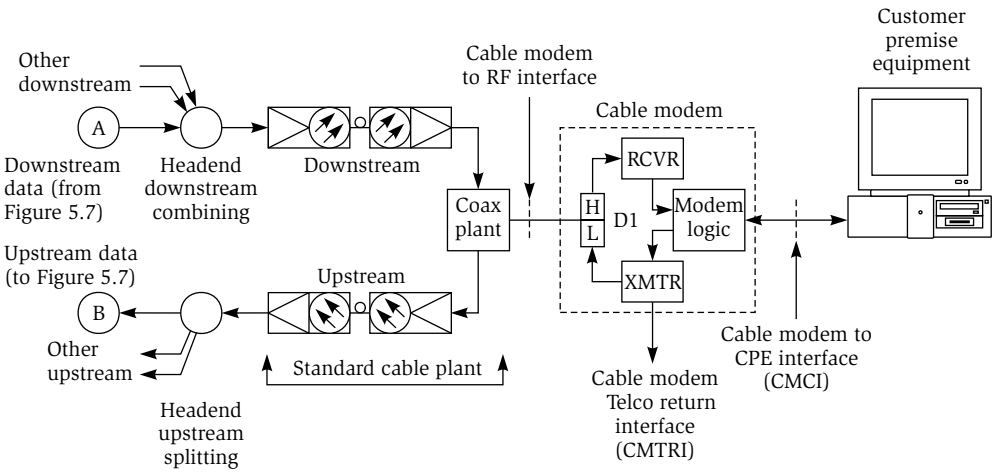


Figure 5.6 DOCSIS reference architecture, headend to home.

direction, the data signals, bubble A, are coupled with other downstream signals and supplied to the downstream HFC network (Chapter 9 of this book covers headend RF architectures in more detail). At the cable modem, the signal is supplied to diplex filter D1, which separates the downstream signal (above 54 MHz) from the upstream signal (below 42 MHz).

The downstream signal is supplied to the receiver in the cable modem, whose output is the downstream data. In turn, that data is supplied to interface logic, which converts the signal to one of several defined interface protocols. One such interface is the 10baseT protocol, described later in this chapter. This protocol is a particular physical layer implementation of Ethernet, which operates at a data rate of 10 Mb/s using four conductors from a CAT-5 cable and RJ-45 connectors (the interconnecting cable looks like a telephone cable with bigger connectors). It has been selected as a standard interface because of its suitable speed and the availability of 10baseT interfaces in many computers. For computers lacking built-in 10baseT interfaces, suitable add-in cards are available.

Two other interfaces are defined for DOCSIS-compliant modems. One is the universal serial bus (USB).[7] The other is the peripheral component interface (PCI).[8] USB has gained acceptance as an interface between computers and a number of external devices. The original USB specification is a 12-Mb/s bus that allows many external devices to be connected to the PC via a hub. The more recent USB specification operates at speed of 480 Mb/s. The PCI interface is used with internal modems and other hardware. The original specification has a maximum transfer rate of 132 MB/s, and is backwards compatible with the older ISA bus. A more recent version of the specification enables transfer rates up to 4.26 Gb/s.[9]

A cable modem transmitter handles return signals bound for the headend. The cable modem acts as an Ethernet bridge at the MAC layer (see later for definition of the MAC layer). This means that it passes traffic from and to the user's computer but does not pass the remainder of the traffic on the cable. Thus, a user doesn't see traffic bound to other users on the node.

An optional telephone return interface is shown in Figure 5.6. Some early systems used telephone return, with data cable-delivered to the subscriber. This allowed modem service on cable systems that had not upgraded to two-way capability.

Figure 5.7 illustrates the headend portion of the reference architecture. Bubbles A and B correspond to the same bubbles in the previous figure. The *cable modem termination system* (CMTS) includes the demodulator and modulator, which interface to the cable system. It also includes the interface to the network. Several options are defined for interfacing with the Internet backbone, either locally or (probably more commonly) by reaching a remote Internet backbone point of presence via a wide area network. The cable modem termination system also interfaces with a security controller and an *operational support system* (OSS).

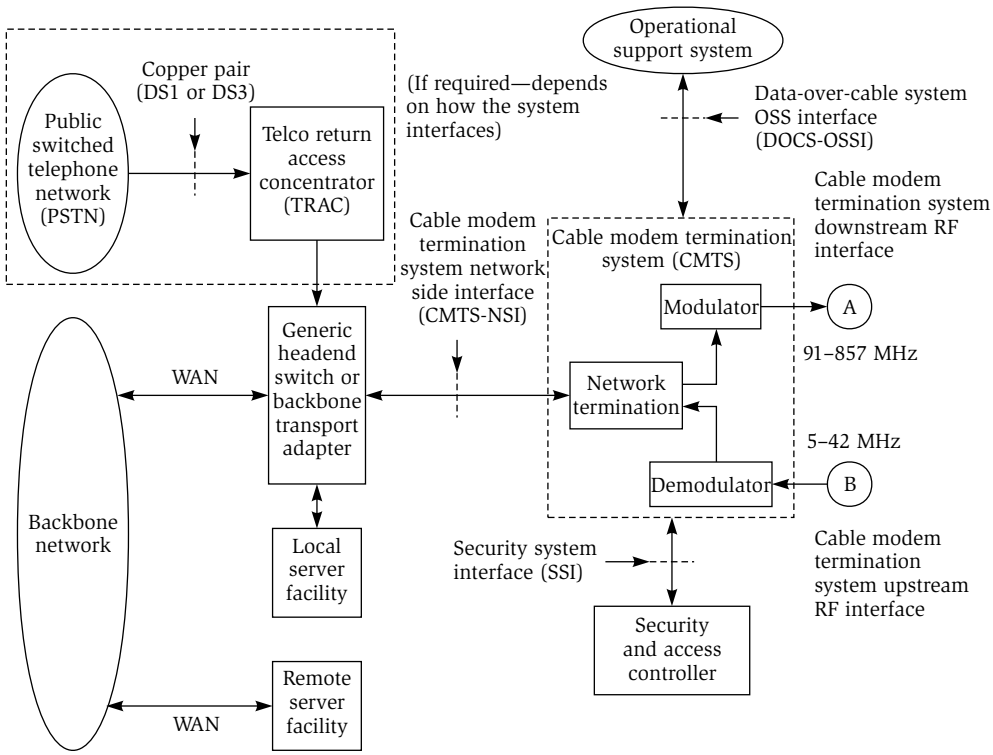


Figure 5.7 DOCSIS reference architecture, headend.

5.2.2 DOCSIS RF Characteristics

Table 5.1 illustrates the RF characteristics of the three versions of the DOCSIS specification. Downstream modulation is the same in all three specifications, being either 64- or 256-QAM, with the characteristics shown. Note that the symbol rate in the 256-QAM mode is higher than it is in the 64-QAM mode but that the excess bandwidth is lower (12% vs. 18%), resulting in a similar occupied bandwidth.

The upstream specification is identical for both DOCSIS 1.0 and 1.1, but in DOCSIS 1.0 the 16-QAM mode was not required and, in fact, was not implemented very often. Thus, the maximum speed of which a DOCSIS 1.0 modem was capable was twice the symbol rate of 2,560 kbaud (thousand symbols per second), equal to 5.12 Mb/s (two bits per symbol in QPSK). This is wire speed, which includes all overhead. In DOCSIS 1.1 the 16-QAM mode is required, so the maximum upstream data rate is twice what it is in DOCSIS 1.0, even though the specifications read identically on this point.

Table 5.1 Comparison of Cable Modem RF Features

Parameter	*DOCSIS 1.0*	*DOCSIS 1.1*	*DOCSIS 2.0*
Downstream			
Frequency (MHz)	91–857 MHz		
Modulation	64-QAM, 256-QAM		
Filter (each end)	$\alpha = 18\%$ root raised cosine (64-QAM) $\alpha = 12\%$ root raised cosine (256-QAM)		
Symbol rates (Mbaud)	5.056941 (64-QAM) 5.360537 (256-QAM)		
Corresponding bit rate (Mb/s)	30.342 (64-QAM) 42.884 (256-QAM)		
Upstream			
Frequency	5–42 MHz (65 MHz maximum, Euro DOCSIS)		
Modulation	QPSK (16-QAM optional, rarely implemented)	QPSK, 16-QAM	Same as DOCSIS 1.1 plus 8-QAM,* 32-QAM,* 64-QAM (S-CDMA and TDMA), 128-QAM (S-CDMA only)
Filter	$\alpha = 25\%$ root raised cosine		
TDMA mode symbol rates (kbaud)	160, 320, 640, 1280, 2560		Same as DOCSIS 1.1 plus 5120
S-CDMA mode symbol rates (kbaud)	Not applicable		1280, 2560, 5120
Channel widths (kHz) @−30 dB	200, 400, 800, 1200, 3200		Same as DOCSIS 1.1 plus 6400
Data rates (kb/s)	320, 640, 1280, 2560, 5120 (QPSK)	Same as DOCSIS 1.0 plus 10,240 (16-QAM)	Same as DOCSIS 1.1 plus up to 30.72 Mb/s in advanced modes
Upstream Preequalization	None	8 taps (opt.)	24 taps

* May not be supported at CMTS.

Early studies of cable modem data indicated a relatively symmetrical data pattern. DOCSIS 2.0 allows a higher bandwidth upstream channel, to approach symmetrical data rates. Two advanced modulation methods were incorporated. One is an advanced TDMA (A-TDMA), with 64-QAM modulation (8-QAM and 32-QAM are optional at the CMTS but required at the modem). Except for the filter excess bandwidth and the use of TDMA, this 64-QAM upstream modulation is identical to 64-QAM downstream modulation. TDMA, you will recall, applies to multiple transmitters, each of which takes some of the upstream transmit time and takes the entire channel; no other modem may transmit at the same time.

The other advanced modulation method is S-CDMA, in which each modem is assigned a spreading code, a pseudorandom sequence used to spread the carrier (see Section 4.3.4). Each modem transmits continuously using its unique spreading code, but at a lower instantaneous speed. Different modems are discriminated by their spreading codes.

A feature introduced in DOCSIS 1.1 and expanded in DOCSIS 2.0 is the use of a preequalizer in the upstream direction. Since the upstream signal is subjected to possible variations in amplitude and group delay as well as microreflections, the CMTS measures impairments and adjusts a tapped filter in the modem to compensate for the impairments. The filter is initially adjusted when the modem is first activated, and is continually updated as the modem responds to the headend. Thus, the filter is termed an *adaptive filter*. Since each modem communicates to the headend through a different path length, and that path is subject to unique distortions, it is necessary to use a different adaptive filter for each modem.

DOCSIS is backward compatible. DOCSIS 1.0 modems will work with a DOCSIS 1.1 or a DOCSIS 2.0 CMTS. DOCSIS 1.1 modems will work with a DOCSIS 2.0 CMTS.

Additional Downstream Physical Layer Features

The downstream physical layer is divided into two sublayers (see Figure 5.2). The parameters described so far reside in the physical media–dependent (PMD) sublayer. On top of the PMD sublayer resides the downstream transmission convergence (DTC) sublayer, which conforms to MPEG-2 (ITU-T Recommendation H.222.0). The DTC sublayer allows for different types of data, such as digital video streams, to be delivered to modems. Virtually any data service desired can be packaged within MPEG packets. The DOCSIS documentation doesn't define these services.

The PMD sublayer, the lowest layer, specifies, in addition to the parameters shown earlier, concatenation of Reed–Soloman block code and Trellis code for error detection and correction. Variable-depth interleaving supports both latency-sensitive and -insensitive data. Section 3.6.2 describes error correction.

Additional Upstream Physical Layer Features

The physical layer in the upstream direction includes other PMD features. The access method is time division multiple access (TDMA — explained in Section 4.1.3), in which a number of modems use the same channel at different times. The standard supports both fixed- and variable-length messages, to improve the efficiency of the return path. Modems may be assigned a time slot, or they may contend for an open time slot. Programmable Reed–Soloman block coding allows trade-offs in error correction and overhead. The standard provides means for controlling the frequency of the modem, the time at which it transmits, and the power level with which it transmits. (Section 16.6.2 contains information related to upstream level control.)

5.2.3 Data Link Layer—MAC

The lowest sublayer of the data link layer is the media access control (MAC) sublayer. The MAC protocol includes the following services:

- Bandwidth allocation, controlled by the cable modem termination system (CMTS).
- A stream of minislots in the upstream direction
- Dynamic mix of contention- and reservation-based upstream transmit
- Extensions provided for future support of ATM (asynchronous transfer mode) or other data PDUs (packet data units)
- Class-of-service support
- Extensions for security as well as virtual LANs at the data link layer
- Support for a wide range of data rates

Bandwidth allocation refers to the process of determining the data rates and frequencies to use in communications. This is done by the modem headend system, the CMTS, based on programmed parameters and existing conditions.

In order to achieve improved efficiency, and acknowledging that many upstream messages are very short, the protocol permits upstream time slots to be subdivided, allowing access by a larger number of modems. Also, data may be sent in packets of variable length, rather than requiring the same length in all packets regardless of the length of the message contained.

5.2.4 Data Link Layer—LLC and Higher Layers

Above the link layer security sublayer, the DOCSIS standard merges with other standards and so is not specified completely in DOCSIS documentation. The logical link control sublayer (class 1 only) is provided in accordance with ISO/IEC 10039. The network layer is the internet protocol (IP), initially version 4,

with an anticipated migration to version 6 as it becomes deployed. Services residing above the network layer use IP as a bearer, and the services are transparent to the cable modem system.

In addition, there are several network management and operation capabilities specified that depend upon the network layer, including the following:

- SNMP (simple network management protocol) for network management
- TFTP (trivial file transfer protocol), a file transfer protocol for downloading software and configuration information
- DHCP (dynamic host configuration protocol), a framework for passing configuration information to hosts on a TCP/IP network
- A security management protocol as defined by DOCSIS

5.2.5 Quality of Service (QoS) and Security

Table 5.2 compares the quality of service (QoS) features and the security of the modem standards. DOCSIS 1.1 and 2.0 are identical in these respects, the difference being in upstream modulation and data rates. The major QoS difference between DOCSIS 1.0 and DOCSIS 1.1 is that DOCSIS 1.0 uses service ID (SID) to identify cable modems and the devices behind them, while DOCSIS 1.1 uses service flows. DOCSIS 1.1 also has improved MAC framing features and improved provisioning and authorization, with advanced BPI+ features.

Quality of Service

Regarding QoS, DOCSIS 1.0 is a best-effort service, meaning that packets are handled when they can be handled. The operator can limit the maximum downstream and upstream data rates and can guarantee a minimum upstream rate. Other than that, there is little the operator can do to control data transmission through a modem. There is no way to tell what the traffic is and no way to limit it or to treat different types of packets differently.

The situation is much different for DOCSIS 1.1/2.0. The concept of *service flows* is introduced as a key element to the operation of the MAC layer. Service flows provide a mechanism for upstream and downstream QoS management. They are integral to bandwidth allocation. A service flow (SF) is a unidirectional datastream for one or more applications, which are to be treated in a prescribed manner during transport. Service flows are described in more detail in Section 5.2.6.

Payload header suppression (PHS) is a technique introduced in DOCSIS 1.1 for improving transmission efficiency, particularly with small packets, such as those used in voice on Internet Protocol (VoIP). The header on each packet can be quite long — on the order of 40 bytes. Compared with a maximum-length

Table 5.2 Comparison of DOCSIS QoS and Security

Parameter	*DOCSIS 1.0*	*DOCSIS 1.1 and DOCSIS 2.0*
Quality of service	Single tier, best effort, 2-way QoS (class of service) profile	Multiple static and dynamic unidirectional service flows with service class names; service-level enforcement statistics
Payload header suppression	No	Yes
Upstream scheduling types	No	BE, UGS, rtPS, UGS/AD, nrtPS
Dynamic service	No	Dynamic service add (DSA), change (DSC), delete (DSD)
Classifiers	No	MAC addr, 802.1Q, P, Ether type, IP addr, IP type, Port #, TOS
Fragmentation	No	Yes
Security	Baseline privacy (BPI)	Baseline privacy plus
Authentication	MAC address checked	X.509 certificate in modem
Key exchange	RSA algorithm with ECB mode of DES for key encryption; keys have a limited lifetime, must be renewed before expiration	X.509 certificates with triple DES; authorization key generated in CMTS, from which traffic encryption keys are derived; sent to modem encrypted with public key; keys have a limited lifetime, must be renewed before expiration
PDU encryption	CBC mode of DES for encrypting PDUs in both directions; keys have limited lifetime	Same as DOCSIS 1.0; multiple security associations for different data flows

Ethernet packet of 1500 bytes, this is not much. But many if not most applications will not be using 1500-byte packets. A relatively slow application would be delayed unacceptably if the transmission system waited until 1500 bytes of data were available. Thus, most packets are much shorter, as short as 64 bytes. In such a small packet, 40 bytes of header almost doubles the length of what must be sent. In order to lessen the overhead burden, DOCSIS 1.1 and later defines payload header suppression.

The header is identical for all packets related to a particular service flow, so for that service flow, the sender suppresses the header, substituting a payload header suppression index, an 8-bit field that identifies the suppressed header (previously agreed between the sender and the receiver). When the packet

reaches the receiver, it looks at the payload suppression index to determine the header that should have been sent and then adds it back to the packet.

Dynamic service will be illustrated in Section 5.2.6. The term refers to the ability of a DOCSIS 1.1 or 2.0 system to provide different services to different classifications of packets. These services may be set up on demand, within the subscriber's authorization. Dynamic services are service *flows* which are created, modified, or deleted after the cable modem registers. They do not force the cable modem to reboot. Compare this to *static service flows* which are defined by a cable modem configuration file installed during registration and after a cable modem reboots. Dynamic service actions include DSA (add), DSC (change), and DSD (delete), which are generically known as DSx.

Upstream scheduling types are also unique to DOCSIS 1.1 or 2.0. These special scheduling types augment the best-effort (BE) scheduling type in DOCSIS 1.0 and are applied to individual upstream service flows by the cable modem under control of the CMTS.

Unsolicited grant services (UGS) allow a modem to transmit or receive packets of data of a fixed size at a regular interval without making a request for each transmission. This is important, for example, in VoIP service, where you cannot tolerate much jitter in the arrival time of packets and the amount of data transmitted at each interval is the same. *Real-time polling services* (rtPS) allow polling of a single modem to see if it has data to send, permitting much more rapid response to a need to send data.

Unsolicited grant service with activity detection (UGS/AD) is a combination of UGS and rtPS suitable for services that must transmit data quickly when it is available, in an environment in which data may not always be available. This would be used by a VoIP application that performs silence suppression. *Non-real-time polling service* (nrtPS) is similar to rtPS, but polling occurs at a lower rate. It is used for such tasks as bulk data transfer not requiring real-time transport.[10]

Classifiers are used to determine the type of packet being handled, for the purpose of assigning it to the correct service flow. Packets may be classified using any combination of header bits. These include the sending and receiving system MAC addresses (a unique permanent address assigned at the time of manufacture of a device) and the IP address (an address unique within at least a network, which may be assigned permanently or on an ad hoc basis). In addition, the DOCSIS system can classify on the universal datagram protocol (UDP) or transmission control protocol (TCP) port number or other characteristics as defined in Table 5.2 and earlier in this chapter.

Fragmentation is useful when handling relatively long packets. A packet may be broken into two or more fragments, which are sent individually, with other data interspersed, and then reassembled at the destination. This allows applications having smaller packets not to have to wait for large packets to be transmitted. This reduces jitter in time-critical applications such as VoIP.

Security

Table 5.2 also shows the security mechanisms built into the modem standards. Security requirements can be broken into several issues. The first issue is authentication. If you are going to send secured data to a modem, you want to know that the receiving device is a genuine modem that will do with the data what you want it to do. You need confidence that it is not a clone (imposter) that may look like a genuine modem but will do nefarious things with the data you send (such as eavesdropping on a neighbor's traffic). In DOCSIS 1.0, authentication consisted of simply getting the modem's MAC address and comparing it to known valid MAC addresses. If it is in the list, the modem is assumed to be good. This really offers no assurance, however, that someone has not built a pirate modem by cloning a legitimate modem's MAC address.

In DOCSIS 1.1, a far stronger authentication mechanism is included, the use of X.509 certificates. These certificates are described in detail in Chapter 22. In brief, the X.509 certificate is a unique electronic certificate stored in each modem and produced by a trusted third party called a certificate authority (CA). The certificate includes information that could have been known only to a legitimate modem, encrypted with the CA's private key. The CMTS knows the CA's public key, which it can use to decrypt the information in the certificate; if the information matches, the modem can be assumed to be legitimate.

Key exchange is the process by which the CMTS and the modem agree on a secret key (the data exchange key) to be used in encrypting data. The process used in DOCSIS 1.0 permitted the data exchange key to be transmitted, albeit encrypted using another encryption algorithm. This is not bad, but if someone were to learn the key exchange encryption key, then the key would be discoverable. In DOCSIS 1.1 and 2.0, the key exchange key is never sent from one side to the other, reducing the possibility that the key exchange key can be discovered.

Packet data unit (PDU) *encryption* is the same in all versions of DOCSIS, but in 1.1 and 2.0 each service flow can have its own encryption key.

5.2.6 More on QoS

Here we discuss the QoS mechanisms of DOCSIS 1.1 and 2.0 in additional detail. Unfortunately, space does not permit a full exposition of the features available, so we shall just provide a feel for what is going on.

Service Flows

Figure 5.8 illustrates several service flows to two different modems. Suppose cable modem 1 is handling two sessions: a voice call and streaming audio. The voice call is critical, in terms of both jitter and delay. The streaming audio requires somewhat more bandwidth. Because it is more tolerant of delay, its decoder can buffer a long string if necessary, making jitter less important.

Thus, the two services are assigned to different service flows. Cable modem 2 is handling a video conference, which requires more bandwidth and is not very tolerant of delay.

Service flows are set up for both modems. Separate service flows are set up in each direction as required, though we show only the downstream direction here. Note that there must be adequate bandwidth in the downstream channel for all service flows, because their bandwidth requirements add. (Figure 5.8 is more of a logical diagram than a physical diagram. Of course, all packets actually enter each modem, and the modem selects only those packets destined for it.)

There are two kinds of service flows: static and dynamic. Static service flows are set up during initial modem registration (when the modem first communicates with the CMTS). At least one service flow in each direction must be established at registration. This is known as a *primary* service flow and can be used for housekeeping and also for best-effort data communications (recall that *best effort* does not imply high-quality communications; quite the contrary, *best effort* means "if and when I can"). Dynamic service flows are established as needed during communications, as sessions are set up and torn down without rebooting the cable modem.

Service flows are identified by a 32-bit number called a *service flow identifier* (SFID) assigned by the CMTS. A service flow may exist even when it is not carrying any traffic. A service flow is said to be *admitted* if the CMTS and possibly the modem are reserving resources. It is said to be *active* if it is actually carrying data. If the service flow is admitted or active it also has a *service identifier* (SID) that identifies the traffic in the upstream direction to the CMTS.

Each packet entering the modem is identified by certain characteristics, described later, and assigned to the appropriate service flow by a *classifier*,

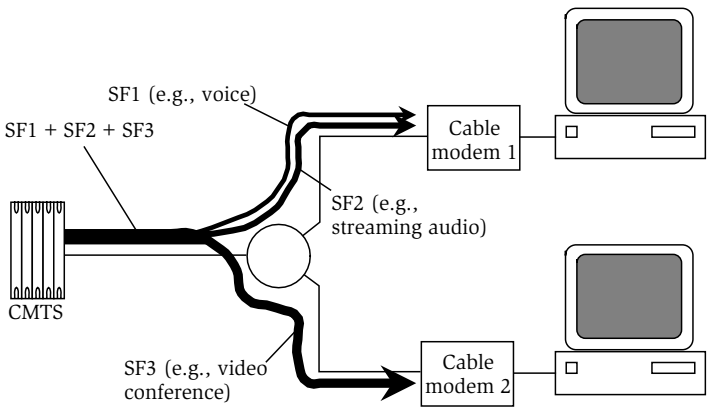


Figure 5.8 Downstream service flows.

a set of criteria used to identify the packet and associate it with a specific service flow. The service flow parameters will determine how to treat the packet. Thus, each service flow is handled in accordance with the policies set up for that service flow. In order to classify the packet, the classifier looks at the header on the packet.

As explained earlier in this chapter, the header comprises a number of bits added to the packet at its beginning. When the packet is placed into (usually) a TCP or UDP (layer 4) packet, a header is added. This header includes source and destination port numbers, which generally identify the program or process using the packet, as explained later in the chapter. When the TCP or UDP packet is placed into an IP packet, a header is added. This header can identify the hardware or portion of hardware using the packet. The link level control (LLC) also adds a header. These headers may be read by the classifier to determine the proper service flow to which the packet should be assigned. The headers can identify the client (end machine or program) sending or receiving the packet as well as identifying the type of packet. From this information, the classifier should have adequate information to identify the service flow, which, in turn, will know how the packet should be handled. If no classification exists for that packet, it is assigned to the primary service flow (the "best-effort" service flow).

An Example Dynamic Service Flow

A dynamic service flow must be set up when needed. To illustrate dynamic service flows, which are set up for one particular session, we use a streaming media example. Suppose the subscriber is downloading a song played in real time or an Internet radio station, something that requires relatively consistent packet delivery. That session is to be assigned its own dynamic service flow. Note in this example that we assume cooperation between the cable operator and the content-on-demand (COD) provider. It also assumes that the client software (on the subscriber's computer) has been customized for this application (or a Web browser is communicating with a COD application on the Web server). If these criteria are not satisfied, the subscriber can still download the streaming media but will have to do it as a best-effort stream, not in a service flow that guarantees the needed bandwidth and jitter.

Figure 5.9 illustrates setup of the streaming media session in a dynamic service flow. The service flow is *dynamic* because it is set up for one session and will be deleted ("torn down") at the end of that session. We are showing here a content server, perhaps owned by the COD vendor, which sends its content upon request to a cache server located in the cable operator's domain.

An optional cache server reduces the number of times the same file must be sent if more than one subscriber wants the same material at different times. The cache server will send the requested file to the appropriate CMTS (through the appropriate routers, using IP to deliver the file to the right point). The CMTS in

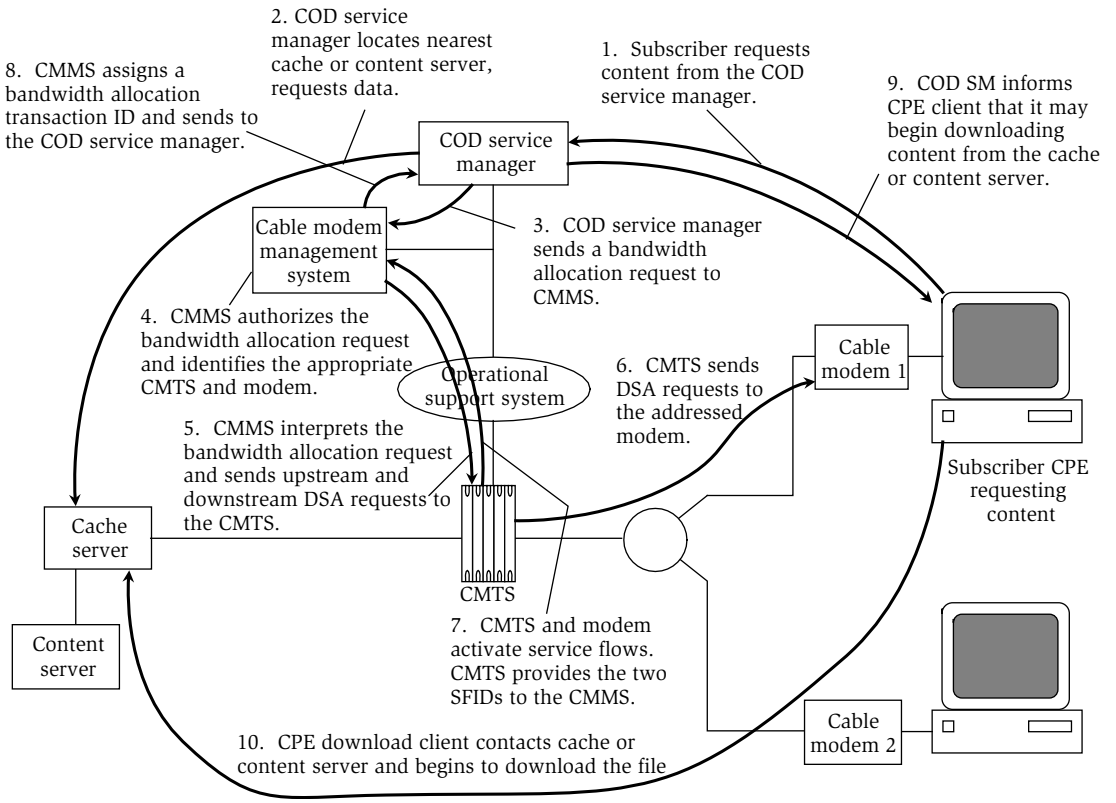

2. COD service manager locates nearest cache or content server, requests data.
1. Subscriber requests content from the COD service manager.
8. CMMS assigns a bandwidth allocation transaction ID and sends to the COD service manager.
9. COD SM informs CPE client that it may begin downloading content from the cache or content server.
COD service manager
Cable modem management system
3. COD service manager sends a bandwidth allocation request to CMMS.
4. CMMS authorizes the bandwidth allocation request and identifies the appropriate CMTS and modem.
Operational support system
6. CMTS sends DSA requests to the addressed modem.
Cable modem 1
5. CMMS interprets the bandwidth allocation request and sends upstream and downstream DSA requests to the CMTS.
Subscriber CPE requesting content
Cache server
CMTS
Content server
7. CMTS and modem activate service flows. CMTS provides the two SFIDs to the CMMS.
Cable modem 2
10. CPE download client contacts cache or content server and begins to download the file

Figure 5.9 Streaming media setup.

turn sends the file to the subscriber's cable modem, which in turn delivers it to the computer.

Also shown are two functions considered part of the operational support system (OSS) for modems in the operator's domain. These are a COD service manager, which contains the relevant policies for offering services such as dynamic service flow, and the cable modem management system (CMMS), which keeps track of what CMTS and modem provide service to which subscriber.

In step 1, the subscriber requests content from the COD service manager, which locates the content on the cache server, if that is where it is, or on the content server. In step 2, the COD service manager requests data from the cache server on behalf of the subscriber (the "client"). In step 3, the COD service manager sends a bandwidth allocation request to the CMMS. In step 4, the CMMS receives the bandwidth allocation request, authorizes it, and identifies

the appropriate CMTS and modem. In step 5, the CMMS interprets the bidirectional bandwidth allocation request and sends a pair of upstream and downstream DSA requests to the proper CMTS, which, in step 6, sends the DSA requests to the addressed cable modem. DSA is one of three types of dynamic service commands used, the other two being dynamic service change (DSC) and dynamic service deletion (DSD). Each of these three commands has a request, a response, and an acknowledgment command. Collectively they are known as DSx transactions. The functions of each are obvious from the name. To add a dynamic service flow, a DSA request is sent, which is acknowledged by a DSA response (usually granting the request), followed by a DSA acknowledgment. Similarly, DSC changes a dynamic service flow and DSD deletes it.

In step 7, the CMTS and modem activate the service flows, and the CMTS provides the two (upstream and downstream) service flow identifiers (SFIDs) to the CMMS. In step 8, the CMMS assigns a bandwidth allocation transaction identification and sends it to the COD service manager. At that point (step 9) the COD service manager informs the CPE client (the software being used by the subscriber) that it may begin downloading content from the cache, if that is where the file is located. Finally, in step 10, the CPE client is now ready to contact the cache server and to begin to download the file. Note that the classifiers installed in the CMTS and the modem for these COD flows recognize the download traffic between the COD cache server and the client and then the selected service flows provide the requested QoS to these packets. Meanwhile, other traffic is managed by the other service flows and is not given the special QoS treatment afforded the COD traffic.

At the conclusion of the session, a similar set of processes is used to delete the service flow. When the COD cache server indicates the download has completed, the COD service manager uses the bandwith allocation transaction ID to inform the CMMS which bandwidth resources to release. The CMMS in turn relays a pair of DSD requests to the CMTS to remove the COD service flows from the modem. Note that if a subscriber attempts to download material using download client software that does not support the request for dynamic service flow, the subscriber can still do so. However, he or she will not receive a dynamic service flow and is at the mercy of how busy the system is to set the rate at which material is downloaded.[11]

5.3 Non-DOCSIS Protocols

This section deals with a number of protocols of importance to cable television technologists. Of necessity, this is no more than a quick-read introduction for those needing basic familiarity with the protocols; it will not make you an expert. Books have been written about each protocol, and we shall reference some of them for those needing more information.

5.3.1 Asynchronous Transfer Mode (ATM)

Asynchronous transfer mode (ATM) is a connection-oriented layer 2 protocol that originated in the telephone industry to provide transport for multiple voice conversations and other data. It features fixed-length cells (as opposed to variable-length packets — see Section 5.3.2 on Ethernet). As shown in Figure 5.10(a), the cells are 53 bytes long, with the first 5 bytes being the destination header. Note that in ATM we refer to fixed-length "cells" rather than the variable-length "packets" of Ethernet and IP. ATM systems operate continuously, with one cell followed immediately by the next, even when there is no data to transmit. Typically, when a connection is set up, sufficient bandwidth is reserved for that connection, and it is reserved until the connection is torn down, regardless of whether it is carrying traffic.

ATM is sometimes used to interconnect hubs serving an HFC plant, but due to cost it is not popular in subscriber service. (Certain concepts of ATM do survive in a number of subscriber-facing services, though.) ATM is known for low jitter (cells arrive at a regular interval) and excellent quality of service. A number of mechanisms are provided to ensure reliable delivery of cells.

As shown in Figure 5.10(b), information delivered to an ATM system interfaces via the *ATM adaptation layer* (AAL), which has two functions. Since many packets from other protocols, such as Ethernet, can be much longer than the 48 data bytes of ATM, these longer packets must be divided into segments at the transmit end and then reassembled into longer packets at the receiver. This is done by the *segmentation and reassembly* (SAR) function.

The other AAL function is performed by the *convergence sublayer* (CS). It provides interface to different types of data packets, called *service data units* (SDUs). Different SDUs require different types of processing. For example, you might be handling video in constant or variable-bit-rate streams. These need a

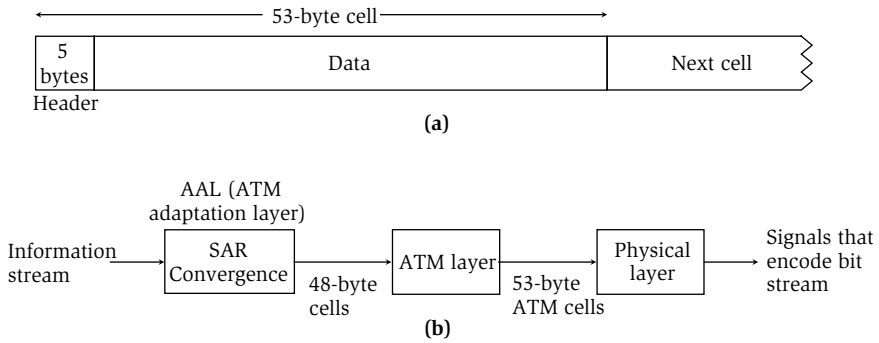


Figure 5.10 (a) ATM cell stream. (b) Building the ATM cell.

different interface, as does an IP packet or other types of data. Five types of CS interfaces are specified.

AAL1 is for traffic generated at a constant bit rate, which must be delivered at that same data rate (after a fixed delay). Intended applications are voice and constant-bit-rate video or audio.

AAL2 is for traffic generated at a variable bit rate. A time stamp is used to control delivery of the packets at the receiver. The SAR must segment variable-length packets generated by the CS. The source and destination must agree on how to deal with timing and cell loss, as in AAL1.

AAL3 is suitable for a connection that occupies one or more variable-sized packets. Two options are provided for error quality. In one option, the CS guarantees error-free delivery. If an error occurs, the destination CS requests retransmission. In the other option, no guarantees are made for error-free delivery.

AAL4 is similar to AAL3 except that it emulates connectionless (datagram) service. The CS layer generates a single variable-sized packet, which is segmented into one or more cells by the SAR sublayer.

AAL5 is for IP packets. The IP packet is packaged into a CS packet that contains a length indicator (for the IP packet) and a CRC calculated over the complete CS packet. The ATM cell header indicates whether or not the cell is the last one of the CS packet.[12]

5.3.2 Ethernet

No layer 2 networking standard is more prevalent today than Ethernet. This connectionless protocol has dominated office networking almost since the beginning of networking, and a large number of homes being built now have wiring in them to enable Ethernet networks. Ethernet was developed at the Xerox Palo Alto Research Center, PARC, by Bob Metcalf in the early 1970s. Metcalf's work built on that of Norman Abramson and his colleagues at the University of Hawaii in the late 1960s. This team developed a network of data communications via radio between various Hawaiian Islands and ships plying the waters around them. Appropriately, the protocol they developed was and is known as *Aloha*.

The pure Aloha protocol was simple: When a station had something to say, it simply transmitted its information and waited for an acknowledgment by the receiving station. If the acknowledgment didn't come in a reasonable length of time, the sending station assumed it didn't get through because another station transmitted at the same time. A *collision* was said to have occurred. It and the other station (which would not have been acknowledged either) waited different random times and then retransmitted.

In any radio communications channel, you cannot assume that a station wishing to transmit can hear another station that is already transmitting, so this

was not required in the Aloha system. Channel utilization is low, because as more stations try to send more data, the probability of a collision increases. Pure Aloha systems can achieve maximum channel utilization of about 18% due to this problem. By coordinating time slots in which stations can transmit, a *slotted Aloha* system is derived, which has a theoretical utilization of about 37%.

In wired networks such as Ethernet, you can assume that one station can hear another, and this allows improvements. Metcalf added a mechanism that listened to a channel before it tried to transmit (*carrier sense*). This should resolve most collisions. But it is still possible for two stations to start transmitting at very nearly the same time, so after a station starts transmitting, it listens to its own outgoing signal to see if a collision occurs (*collision detection*). When a collision is detected, the station stops transmitting and waits a random time before trying to transmit again. Multiple stations can access this network, so we have *carrier sense, multiple access with collision detection*, or a CSMA/CD system.

The Ethernet standard was published in 1980 by a consortium of three companies, Digital Equipment Corporation (DEC), Intel, and IBM. This standard has become known as the DIX standard, after the three companies. About the same time, the IEEE[13] started to document the standard, and continues today to be the custodian of the Ethernet standard, known as IEEE 802.3.* (There are other communications standards in the 802 family — perhaps one of the best known today is the wireless 802.11 set of standards.) The original DIX standard (which reached version 2.0) and the 802.3 standard are similar, though there are a few differences. We shall describe the IEEE version of Ethernet.

There have been a number of different physical layer implementations of Ethernet; here we discuss only the most commonly encountered implementations. Ethernet physical interfaces are described by a number indicating the speed, followed by the word *Base*, followed by a designation indicating the medium over which the signal is transmitted. 10BaseT indicates a 10-Mb/s transmission speed over twisted pair cable. 100BaseTX is a 100-Mb/s signal transmitted over the same types of twisted pair cable. These two form the basis for the most common Ethernet interfaces in use today. Much equipment can interoperate between the two speeds; this is known as a 10/100BaseT interface. Other popular forms of Ethernet operate at a date speed of 1 Gb/s: 1000BaseT (twisted pair), 1000BaseSX (short-distance optical using multimode fiber), and 1000BaseLX (long-distance optical using single-mode fiber). 10-GB Ethernet has been standardized.

10/100BaseT

There is almost zero probability that your office is *not* using 10/100BaseT for its internal network. The numerals before the word *Base* represent the signaling

---

* The 802.3 standard does not use the term *Ethernet*, because the IEEE is properly concerned about being tied to any commercial terms.

speed, either 10 or 100 Mb/s. The *T* indicates twisted-pair transmission medium. Technically, there should be a dash between *Base* and *T*, but we follow the common practice of omitting the dash. There are several variations of 100BaseT communications. The one in common use is really 100BaseTX, but since the others are not used, often the *X* is omitted.

Many homes built in this century are wired with cable to support 10/100BaseT data transmission. The medium is two twisted pairs, one for each direction of communication. The gage of wire, the insulation, and the twist are selected so that the differential characteristic impedance is 100 ohms. Characteristic impedance has exactly the same definition as in the 75-ohm coaxial cable you are used to, except that the transmission is balanced with respect to ground rather than single-ended.

Most of the cable used today is unshielded twisted pair, designated UTP. You may on rare occasion see shielded twisted pair STP. The standard cable consists of four twisted pairs, of which two are used for one Ethernet connection. There are varieties of Ethernet that use all four pairs, but you are not likely to encounter them until you get to gigabit Ethernet. While emphatically *not* recommended, sometimes the two other twisted pairs are used for telephone connections. Due to crosstalk potential, it is *strongly* recommended to use separate cables for voice and data, in both residential and commercial wiring. The connector is the RJ-45, which looks like a larger version of the RJ-11 connector used for telephones. (RJ stands for *registered jack*, a U.S. telephone industry term.)

Manchester encoding is used to provide a clock signal and to remove any dc component of the data. Manchester encoding is described in Section 4.2.2. The voltage on each pair is balanced, with a minimum amplitude at the transmitter of 900 mV pin to pin (p–p) and a maximum voltage of 2.63 V p–p at the crest of an overshoot. The receiver works with as little as 160 mV p–p. Contrast this with the ringing voltage of a telephone pair, which can be as high as 108 V, and you see why it is not a good idea to run telephone in the same cable as Ethernet. Unfortunately, some home builders are wiring both phone and Ethernet connections in the same cable, despite an abundance of wisdom that says not to do it.[14] Laboratory tests done by the writer in 2003 indicated that it will sometimes work, but in no way can we endorse this reckless practice.

1000BaseT

Gigabit Ethernet was standardized in 1999. There is one standard for twisted-pair cable and two for optical cable. 1000BaseT is the reference for the twisted pair standard, which uses all four conductors of a Cat5 or better cable (see later). The signal encoding, called 4D-PAM5, consists of four symbols, each of which contains two bits. All four wire pair are used to transmit signals in each direction, so 8 bits are transmitted at a time. A *hybrid* is used to separate transmission in each direction, similar to what is done in conventional analog telephony. Hybrids may be built with transformers or with operational amplifiers. As with the slower speeds, balanced transmission is used, with a voltage range of ± 1 V p–p.

1000Base-X (Fiber Optics)

There are two forms of gigabit Ethernet over fiber-optic cable in use. 1000Base-LX is a long-wavelength (usually 1310 or 1550 nm) signal using single-mode fiber just as is used in HFC plant. It is intended for long-distance operation, with commercially available transceivers (often called *GBICs*) for distances exceeding 70 km. 100Base-SX is a short-wavelength, short-distance transmission standard that uses multimode fiber. Both standards use one fiber for each direction.

The signal uses 8B/10B encoding, in which a 10-bit symbol is substituted for every 8 bits of data. This does create 25% overhead — the actual transmission rate for gigabit Ethernet is 1250 Mb/s. It is used to ensure adequate transitions for clock recovery and no dc component. 8B/10B encoding is explained more fully in Chapter 19.

Twisted-Pair Wiring

The cable commonly used is called category 5, or simply *Cat5* cable. There have been several generations of cabling used for voice and data, with the number incremented for cables that exhibit better return loss and radiation characteristics. Besides Cat5, there is a Cat5e (*enhanced*) cable and a Cat6.

Figure 5.11 illustrates two standard ways to wire RJ-45 plugs. The Telephone Industries Association (TIA) standard T568 lists two wiring sequences, both shown in Figure 5.11. The "preferred" T568A standard was formerly used more often, but today the optional T568B wiring sequence seems to be more popular. The only difference is that wire pairs 2 and 3 are interchanged. Note that the wire pair numbering always stays with the same color of wire pair. Pairs 1 and 4, which are not used in the most common 10/100BaseT wiring, are the

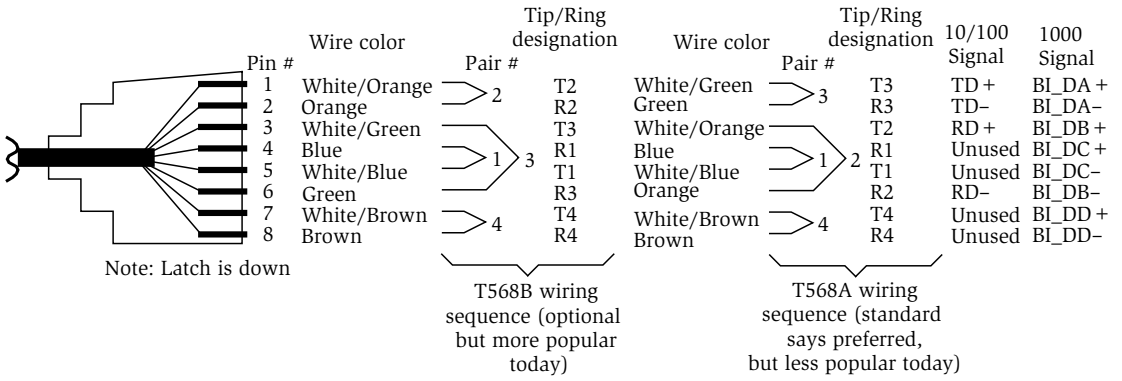


Figure 5.11 Common RJ-45 wiring.

same for both options, but wire pairs 2 and 3, which are used for the most common Ethernet wiring, are different in the two options.

The "signal" designations are for transmit data + and − and receive data + and −. For gigabit Ethernet on twisted pair (1000BaseT), different designations (standing for *bidirectional*) are used, since all pairs are used and signaling is transmitted in both directions simultaneously. Recall that these are balanced transmission lines, so the designation is to allow correct polarization, or phasing. The *tip/ring* designations go back to the early days of telephony, when 1/4-inch phone jacks, similar to the large jacks used in some audio equipment and for headphone connectors, were used in central offices (COs). The connection on the end of the plug was the "tip" and the other connection was the "ring." Note that the ring wire is always a solid color, and the tip is always white with a small trace of that same color.[15,16] (For stereo audio connections made on 1/4-inch phone jacks, a third connection is added called a *sleeve*.)

The important thing is that the two ends must be wired pin to pin on each end, and the twisted pairs must be followed. If you violate the twisted pair, for example, by interchanging the wires on pins 2 and 3 so that you still have pin-to-pin electrical continuity but don't have the transmit and receive pairs as separate twisted pairs, then the cable will probably not work, because you have upset the 100-ohm characteristic impedance, setting up standing waves (in the baseband data world this will manifest itself as pulse distortion). Furthermore, you will have excessive crosstalk between transmit and receive as a result of not having one signal (transmit or receive) on a twisted pair. A twisted pair driven with a balanced source and load does not radiate or pick up radiation, at least to the extent the system is truly balanced. This is the same as the issue described under Line Balance in Chapter 6.

Interconnecting Ethernet Devices

Figure 5.12 illustrates a basic in-home (or office) network with two computers and a game console. Other devices could be connected. All wiring is home-run from a hub or central device. The specified maximum distance between any endpoint and the hub is 100 meters. The end devices, computers, game consoles, or whatever are known as *data terminal equipment* (DTE). The central device has a number of names, depending on what it does and who is naming it.

Ethernet connections (at least the common ones you are likely to encounter) are never looped through from one device to another. Despite the attempts of some home installers to install loop-through connections, it doesn't work. *Loop-through* means that the same wire pairs are run from one outlet to another, after originating at a hub. Suppose you have a loop-through connection as shown on the game console in the figure. If the game console (or computer or whatever) is placed at the end of the loop-through, it will work. But if it is placed at the first point, as shown, the cable from there to the end of the loop-through will act as an unterminated stub and will reflect pulses, disrupting the performance of the

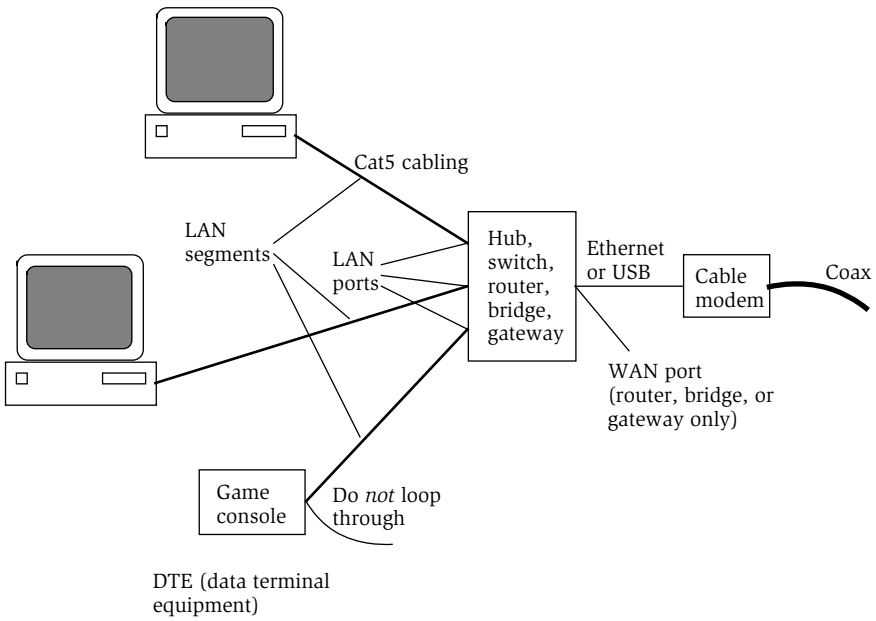


Figure 5.12 In-home network.

link. We have found that in some cases, terminating the end connection point will allow the system to work, but this is only an emergency solution, not a recommendation.

It is true that one Ethernet connection will listen for others transmitting before it transmits, but physically what happens is that there is a central hub that can echo signals entering one port to all other ports. The hub may perform other functions as well. A *hub* is a common connection point for devices in a network. Hubs are commonly used to connect segments of a LAN, as shown in Figure 5.12. When a packet arrives at one port, it is copied to the other ports so that all segments of the LAN can see all packets.

A *switch*, or sometimes *switching hub*, is a bit more intelligent. It filters and forwards packets between LAN segments. Each DTE is assigned a unique identification called a *MAC address*, described later. A switch learns the MAC address of each DTE connected to it; when it gets a packet with that MAC destination address, it forwards the packet only to the appropriate port. The switch learns the MAC address of each DTE by listening to what comes in on that port. Each transmission includes a destination MAC address and a source MAC address. The switch listens to the source addresses to learn what is on each port. This usually suffices to allow the switch to know how to route any packet it receives.

However, sometimes a packet will arrive with a destination MAC address the switch doesn't recognize. In that case, the switch will replicate the packet on all ports, just as a hub would do. When the correct port (if any) receives the packet and sends an acknowledgment, the switch will learn the correct port for that MAC address.

A *router* is a yet more intelligent device that can operate at layer 3 as well as layer 2. It often is placed between two LANs or between a LAN and a WAN. We shall cover their functions in more detail later.

A *gateway* is a router used to connect two or more networks, as shown in Figure 5.12. It may have a number of higher-layer services, which will be described below.

A *bridge* is a low-end gateway that interconnects two LANs or two segments of the same LAN.

Ethernet Frames

Figure 5.13 shows the makeup of an Ethernet *frame*. A frame holds one *packet* of data. The first 8 bytes are the preamble. Some Ethernet systems don't transmit continuously, so the preamble is used to synchronize a receive clock before data is transmitted. In other Ethernet systems, data transmission is continuous, but the preamble has been retained to keep the frame structure.

The next 6 bytes are the destination address — the MAC address of the intended recipient of the packet. Next is the source address, the MAC address of the sender. This is needed to allow the recipient to respond to the source and also to allow switches to learn the MAC address on each port, for efficient packet routing. Following the address fields is the type/length field, which is used for several purposes. It can identify the higher-layer network protocol carried in the data field, as described next. It also identifies the length of the packet. Unlike ATM cells, Ethernet packets are of variable length, so something must tell the receiver when a packet ends.

Next is the data, the thing we want to transmit. It may have headers for higher-layer protocols embedded in it. The length of the data is between 46 and 1,500 bytes. Following the data is the frame check sequence, a cyclical

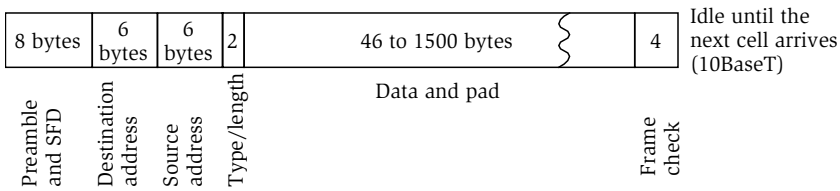


Figure 5.13 Ethernet frame.

redundancy code used for error checking but not correction. Error checking is described briefly in Chapter 3.

Ethernet Addressing

Every Ethernet port in every piece of equipment is assigned its own unique address, called the MAC (media access control, a layer 2 feature) address. Each manufacturer applies for and receives a range of MAC addresses, which are assigned individually to each and every piece of hardware manufactured. The MAC address is 48 bits (8 bytes) long, so the total available address space is about $2^{48} = 281.5 \times 10^{12}$ addresses. The first 24 of those bits are called the *organizationally unique identifier* (OUI). They identify the organization that builds the interface. The IEEE Standards Association administers this portion of the address field.

This assigned address is a *unicast* address, meaning that it is used when another device wants to contact that Ethernet port and only that Ethernet port. When a device starts to receive a new packet, it listens to the destination MAC address. If the transmitted address matches its address, it receives the packet. Otherwise, it discards it. However, there are times when you need to communicate with multiple devices. You can use *multicast* addresses for this. Any number of Ethernet ports may be set to listen to the same multicast address and to receive the data. Finally, there is a *broadcast* address (all 48 bits are 1s), used to communicate with all devices on the network.[17]

Ethernet addresses are written in a manner intended to make it easy to deal with a 48-bit binary number. They are written as pairs of hexadecimal numbers, separated by dashes. A *hexadecimal* number is a base-16 number represented as shown in Table 5.3. Each byte (8 bits) of the MAC address is written as two hexadecimal numbers, separated by a dash. For example, the Ethernet card in the writer's computer, which interfaces to the in-home network, has the MAC address 00–03–6D-15-C0–97. Note that there are six sets of two hexadecimal numerals, each set representing 8 bits, for a total of 48 bits.

The card that handles the interface between the computer and the network is called the *network interface card* (NIC). Even if the network interface is built into the computer or other equipment, it is still called an NIC.

5.3.3 Internet Protocol (IP)

Internet Protocol (IP) is not just the Internet. It is a very important connectionless layer 3 protocol used to move packets around on the Internet as well as on a lot of other networks. IP allows any computer on the Internet (or another IP network) to locate and exchange packets with any other computer. You often see TCP/IP together, where TCP is a layer 4 protocol often used with IP (as in, for instance, UDP). TCP and UDP are discussed later.

Table 5.3 Hexadecimal Numbering

Base 10	*Base 2*	*Hex*
0	0000	0
1	0001	1
2	0010	2
3	0011	3
4	0100	4
5	0101	5
6	0110	6
7	0111	7
8	1000	8
9	1001	9
10	1010	A
11	1011	B
12	1100	C
13	1101	D
14	1110	E
15	1111	F

IP addresses are expressed in the *dotted decimal* format. In version 4 of IP, addresses are 32 bits long. Table 5.4 illustrates the dotted decimal address. The top row gives the dotted decimal representation; the second row expresses this as individual decimal numbers, the third row shows the corresponding binary numbers. This IP address was assigned to the WAN port on the writer's gateway connected to his cable modem at the time of writing. The assignment came from his cable operator. As shown later, the IP address may change from time to time.

The contents of an IP packet are called a *datagram*. Each datagram, as it traverses the Internet, retains the same source and destination IP addresses, even as it traverses different layer 2 networks. For example, a packet may travel part of the distance on a token ring network, partway on ATM, and partway on Ethernet. These are layer 2 protocols, and each will put on its own layer 2 header as the IP datagram enters that network and strip its layer 2 header when it is handed to another network. But the IP header stays with the datagram during the whole process (though it will be modified in each router through which it passes), defining the ultimate destination of the datagram. (An exception is that

Table 5.4 Dotted Decimal Notation for IPv4 Addresses

Dotted decimal	*24.98.160.78*			
Base 10	24	98	160	78
Base 2	00011000	01100010	10100000	1001110

a firewall between networks or between a network and a computer may substitute its own IP address for that of the ultimate process using the datagram.)

When switching in a network is done by looking at the layer 3 (IP) address, the device doing the switching is usually called a *router*. Routers technically have hosts built in, because they themselves can be addressed (they have their own IP addresses), for management purposes.

IPv4 Versus IPv6

IP version 4 (IPv4) is in common use today. For years there has been a desire to replace IPv4 with IPv6, but the difficulty of managing such a change has prevented widespread deployment of IPv6. There are several new features in IPv6, the most notable being the expansion from a 32-bit address to a 128-bit address. This will simplify address management, removing the need for assigning temporary private IP addresses to devices. In addition, the header format is streamlined. A flow label is added to denote a stream of traffic with special QoS needs. Support for both authentication and confidentiality are mandatory, and there are a few other changes. Unless otherwise noted, we are describing IPv4 in this book.

IP Embedded in Ethernet

Figure 5.14 illustrates an IP datagram inside an Ethernet frame. This figure is laid out with four bytes (32 bits) in a horizontal row. Each horizontal row follows serially the one above it.

Compare Figure 5.14 with the illustration of an Ethernet frame in Figure 5.13, which identified all the elements of the Ethernet frame consistent with Figure 5.14. The "Data and pad" of Figure 5.13 are expanded in Figure 5.14 to show the elements of the IP datagram. The elements of the IP header are summarized next.

The IP Header

The first four bits, *vers*, contain the IP version number, currently 4 (0100) and possibly going to 6 (1010) at some point. The next four bits, *hlen*, counts the number of 32-bit units in the IP header. The basic IP header as shown is 20 bytes, or five units of 32 bits. We show later that some IP headers are longer as a result

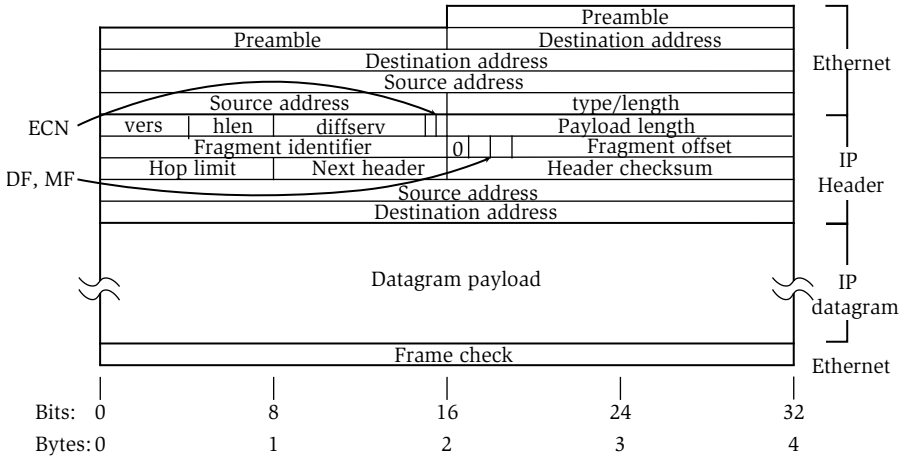


Figure 5.14 IP Datagram Inside Ethernet Frame.

of options being embedded. IP defines a number of options that may create a longer header. These are covered in the endnotes. Options include, but are not limited to, the ability to record the route taken by the datagram, the ability to force a particular route, and time stamping.

The next field is a 6-bit field labeled *diffserv*, for differentiated service. The 6 bits contain a so-called *diffserv codepoint* (DSCP) value, which tells a router how to handle the packet. Special treatment may include expedited forwarding or extra protection from loss. Most values for this field are not defined in IP but rather are used by different networks in accordance with each network's policies. This field is an important enhancement to IP to bring it closer to parity with the QoS capabilities of ATM or similar protocols. It allows datagrams to be given higher priority if the datagram is something, such as voice, that needs to get through a network quickly. (Note, however, that just because ATM has good QoS features does not mean that they are always used.)

The final two bits in this field, ECN, are the *explicit congestion notification field*, which provide cooperating routers with a way to improve management of congestion on a link. When a router handling the packet notices that a link is getting congested, it sets these two bits to 11 before forwarding the packet on the congested link. When the packet is ultimately received, the receiving host sees that one of the routes the packet took is congested, so it slows down the sender, using methods shown later in the section on Transmission Control Protocol.

The next two bytes are the *payload length*, which defines the length of the datagram, not counting the header. IP datagrams can be quite long, though they may well have to be segmented into several packets or cells (depending on the protocol on which it is carried) on any particular transport segment. For

example, if the datagram is being carried on Ethernet, as shown in Figure 5.14, the maximum amount of data in the Ethernet frame is 1500 bytes (this includes the IP header and the datagram). If the datagram goes through an ATM network, it will have to be segmented into cells no longer than 48 bytes, as described earlier. If the datagram is too long to fit, it will have to be segmented into several packets. The *fragment identifier* field is added at the sender, to be used to uniquely identify the packet to which a fragment belongs so that the fragments can be reassembled at the destination. This field is set to a value even if the sender is not fragmenting the packet, because it is possible that a later network will have to fragment it.

The next three bits are flags that instruct routers on segmentation. The first bit is always 0. The next bit, DF, indicates that the packet may not be fragmented. If a router encounters a datagram that is too long and the DF bit is set, then the router has no recourse but to discard the packet. The next bit, MF, indicates that more fragments follow. It is set except on the last segment of a fragmented datagram. Finally, the *fragment offset* field specifies how far into the datagram the fragment is so that the fragments can be reassembled as one datagram by the time they reach their destination (if not before). It is quite possible that the segments will arrive out of order, so segments received before predecessor segments will have to be held until all segments are received.

The next field is the *hop limit*. (This was originally called the *time to live* (TTL) field). The sender sets this field to some value; every time the datagram passes through a router, the field is decremented by 1. If the value reaches 0, the packet is discarded. The hop limit protects the Internet against an undeliverable packet, which would otherwise hop around a network forever if it could not be delivered. This can happen if a *routing loop* is created; that is, an error in a routing table sends a packet back to the previous router that handled it, which sends it to the router with the errored routing table, which sends it back, etc. If nothing caused the packet to be discarded, it could bounce around forever in the Internet, tying up resources needed by packets that can be delivered.

Next is the *next header* field, also called the *protocol field*. This indicates the next higher header, which may be a layer 4 header or may be intimately related to IP. We shall discuss some of these protocols later. Finally, the *header checksum* provides protection against corruption of the header. Each router that handles the datagram computes the checksum by doing a 16-bit addition of the header, discarding carries. The receiving router does the same computation; if the two disagree, it means that an error was made in the header, so the entire packet is discarded.

Finally, the IP header includes the source and destination IP addresses. These don't change as the datagram traverses the network. Each router handling the packet reads the destination address to know where to send the packet (controlled by a massive and continuously changing routing table in the router) and reads the source address to update its routing table for that address.

Types of IP Addresses

There are three types of IP addresses: unicast, multicast, and broadcast. *Unicast* addresses are the most straightforward. They are assigned to identify a single interface (for most practical purposes this equates to a single host, though there is a difference). A unicast message is intended for one and only one interface. *Multicast* addresses identify a set of interfaces. This allows a message to be generated once and sent to a number of interfaces. A router with receiving hosts attached to two or more ports will replicate the packets on each port. Multicast addressing is useful for video conferencing, in which the same signal is to be sent to several participants. Similarly, it is used in IP distribution of video (IPTV). If one router needs to pass messages to several other routers, it can do so using multicast addresses. *Broadcast* addresses are a special case of multicast addresses. They identify all interfaces on a network. The use of broadcast addresses is discouraged.

Internet Control Message Protocol (ICMP)

Internet Control Message Protocol (ICMP) is a layer 3 (network) protocol, but it sends its messages inside IP datagrams. The *next header* field (see Figure 5.14) identifies ICMP. ICMP coordinates many aspects of the operation of a network, including controlling address assignments, reporting errors, and providing diagnostic support.

Router discovery is one important function it performs. This allows hosts (any device on a network) to discover what routers are connected to the network. Periodically, routers send *router advertisement* messages to all systems on the network. The router advertisement message includes the address of every router known to the sending router and a relative preference value. When possible, all hosts should select the router with the highest preference value. The message also includes a lifetime field. If a host doesn't hear from a router within that lifetime (typically 30 minutes), it must assume that the router is no longer available. The router advertisement messages are sent more frequently than this.

Just knowing the addresses of all routers on a network does not ensure that a host will direct a message to the router that is the best path to the destination. It selects a router to which to send the message. If it selects the wrong router, that router redirects the message to the correct router and also sends an *ICMP redirect* message to the host, saying, in essence, "Next time you want to send a message to this host, send it to this other router, not to me."

Internet Group Management Protocol (IGMP)

Like ICMP, *Internet Group Management Protocol* (IGMP) is a protocol in its own right, but it is an integral part of any IP implementation. Hosts use IGMP to announce (and later renounce) their membership in groups. Routers listen to

these messages to track group membership. They then know how to forward datagrams addressed to groups.[18]

Dynamic Host Configuration Protocol (DHCP)

IP addresses are assigned to individual hosts, that is, all computers on a network. These IP addresses may be taken from public IP addresses, or, since IPv4 has a limited address space, most machines that don't have need for a public address are assigned private IP addresses owned by the network to which the host is attached. This allows reusing addresses on the Internet and simplifies the management of the network. IP addresses may be assigned manually to hosts, or they may be assigned automatically. Automatic assignment is usually preferable when possible.

The protocol used to assign IP addresses is the *Dynamic Host Configuration Protocol* (DHCP). When a host, such as your PC, comes up on a network, it asks for an IP address if it has not been assigned a permanent one. The host sends a DHCPDISCOVER packet. Connected somewhere to the network is at least one *DHCP server*, whose job it is to assign IP addresses to hosts. When a DHCP server receives a DHCPDISCOVER packet, it returns a DHCPOFFER packet containing an IP address and other information. It is possible for more than one DHCP server to be resident on a network, in which case the host receives more than one offer. It selects one and sends a DCHPREQUEST message to that server. Assuming the offer is still valid (for example, the IP address has not been assigned to another host in the meantime), the DHCP server returns a DHCPACK (DHCP acknowledgment) message.

Once the host receives the DHCPACK message, that IP address belongs to it for the duration of the *lease*, that is, the time for which the DHCP server has said that the address is valid. Before the end of a lease, the host must request another IP address by starting the process again. The lease allows an IP address to be placed back in the pool of available addresses if the host goes away for any reason, such as a power-down.[19]

5.3.4 Layer 4 Transport Protocols

Most Internet traffic is handled using one of two layer 4 protocols: TCP or UDP. Layer 4 protocols have two primary responsibilities. First, they must distinguish between different applications running on a system. On the same computer, you can download email, stream audio, and open a Web page all at the same time, over the same physical connection. It is the layer 4 protocols that are making sure packets are directed to the correct application.

The second responsibility of layer 4 protocols is reliability. IP is a best-effort service, which does not guarantee delivery of packets and does not guarantee that the packets delivered will be error free. Layer 4 protocols either improve the reliability of delivery slightly (UDP) or they improve it greatly (TCP).

Ports

In order to distinguish between applications, the layer 4 protocol uses a *port number*. Data is transferred between a communications system and an application using a port number. A good way to think of a port is to recognize that every application getting data from a network must somehow transfer that data in the computer, from the program that sends and receives data to the application program. For example, suppose you are streaming audio. IP is running in your computer to receive datagrams bearing the streaming audio. Another application is running to play the audio. The two must exchange data. The application playing the audio asks for a port number when it starts. The operating system assigns a memory address corresponding to that port number. Data from the communications program is placed in that address and then picked up by the streaming audio program. The address performs the function of a mailbox.

When an application requests data, for example, by your clicking a hot key for a streaming audio site, the application sends a request to the server having the data, and in the request messages it includes the port number for the return datagram. The server puts this port number in the return datagram. The communications software was informed by the operating system of the address corresponding to the port number, so when it gets a datagram with that port number, it knows where to store the datagram in memory. The streaming software then reads that address and processes the datagram into audio.

Some port numbers are created on an ad hoc basis as they are needed; others are standardized and predictable. The standardized port numbers are called *well-known ports*. Table 5.5 lists a few of the well-known ports.

Sockets

Ports identify application protocols, but they don't identify application programs running in real machines. To do so requires not only a port number but also the address of the machine running the program. This combination of a port number and an address is known as a *socket*.

Table 5.5 Examples of Well-Known Ports

Port(s)	*Transport(s)*	*Application*
20, 21	TCP	File Transfer Protocol: used to transfer files between computers
25	TCP	Simple Mail Transfer Protocol: for email transfer
53	TCP, UDP	Domain name service: maps a URL to a server's IP address

User Datagram Protocol (UDP)

User Datagram Protocol (UDP) is a layer 4 protocol used to deliver datagrams not requiring or not being able to use some of the features of TCP, discussed shortly. UDP is a simple, connectionless protocol that is the layer 4 protocol of choice for such applications as streaming audio and video and VoIP (voice on Internet Protocol — see Chapter 6) packet delivery. Applications using UDP are frequently characterized by the need for rapid (low-latency) delivery, where it is not feasible to correct for transmission errors. UDP is also used where a transmission is bound for more than one receiver, making acknowledgment moot.

For example, suppose you are handling a VoIP voice transmission. If a packet is received in error, there is not much you can do about it anyway. You could ask for retransmission, but by the time the request goes to the sending machine and it retransmits the packet, it is much too late to use that packet. So you do the best you can with the packet received in error. This might mean that you discard it and replace it with the previously received good packet. Similarly, you don't have the luxury of reordering packets received out of order, because this will require time, and time is of the essence in handling audio packets. Two other protocols often used with UDP are the *Real-Time Protocol* (RTP) and *Real-Time Control Protocol* (RTCP), both described in Chapter 6.

UDP headers and data reside in the datagram payload of Figure 5.14. The UDP header is simple, consisting of just four parts: the source port, destination port, datagram length, and checksum. Following the header is the application data, the desired information to be transmitted.

Transmission Control Protocol (TCP)

Transmission Control Protocol (TCP) is a more sophisticated, connection-oriented protocol used for reliable delivery of data that is not required to be delivered in real time. TCP can correct errors in transmission. It can detect packets received out of order and put them back in the correct order. It eliminates duplication of packets and requests missing packets.

Figure 5.15 illustrates a TCP message in an IP datagram, in an Ethernet packet. Remember that the Ethernet encapsulation may or may not be present, depending on what layer 2 network is carrying the datagram. The first two elements of the header are the source and destination ports, the purpose of which was explained earlier.

The *sequence number* is the field that allows packets received out of order to be put back in the correct order. The sender assigns a sequential number to each byte of the application data. The sequence number is the sequential number of the first byte of the TCP packet. When the receiving application receives a packet, it looks at the sequence number and compares it with the expected sequence number, computed by adding the number of bytes in the previous packet to the sequence number of that previous packet. If the sequence number is the one expected, then the receiver knows that it can simply assemble the data

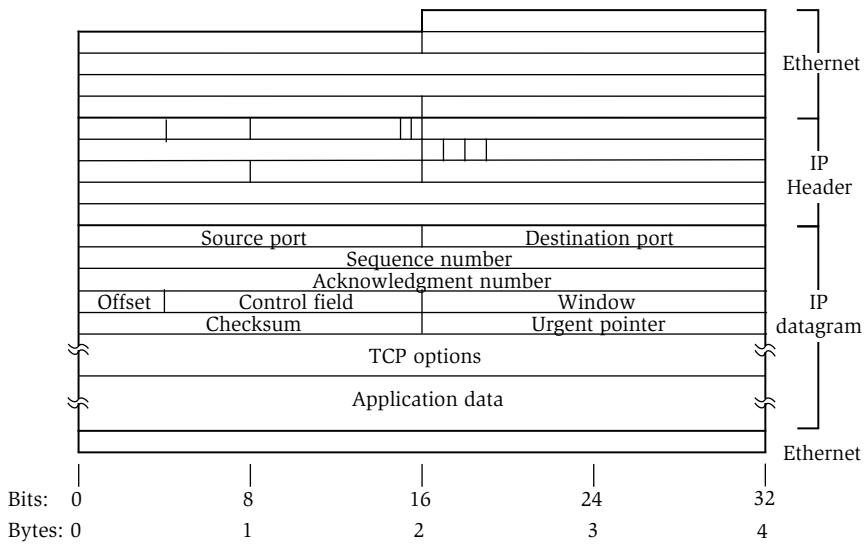


Figure 5.15 TCP in IP in Ethernet.

in the two packets to pass it to the next higher layer. If a sequence number is duplicated, it means that a packet has been duplicated, so the extra copies can be discarded. If there is a gap, it means that a packet is missing. It may come later, in which case it is inserted in its rightful place. Or it may not come, in which case the sender must retransmit it. The sequence number also plays a role in flow control, a process whereby the receiver can limit the amount of data sent to it.

Since TCP is a connection-oriented protocol it must acknowledge receipt of data. This is the purpose of the *acknowledgment number* field, which is used by the receiver to acknowledge the packets received from the sender. It has meaning only when the ACK (acknowledgment) bit in the control field (see next paragraph) is set, which is the usual case when TCP is transferring information.

Next is the *offset* field, which tells the receiver where in the packet the data starts. Since there can be options in the TCP header, the receiver must know where to start looking for data. The *control* field consists of 12 single-bit flags, of which the first four have no meaning at present. The remaining flags signal such things as congestion status, use of the acknowledgment and urgent fields, sequence number synchronization, and a flag to indicate that the sender is finished sending.

The next field, the *window*, indicates how much data the receiving system is capable of accepting. This is a critical flow-control mechanism. It is quite conceivable that a machine receiving data will not be able to process it as fast as the sender can send it. For example, suppose a file server at the headend is downloading audio files to a small device, such as an MP3 player. Over a fast

connection, it is quite possible that the server can get data to the MP3 player faster than it can store it in memory. The MP3 player must have some way to say, "Slow down." It decreases the size of its window when it acknowledges a packet, telling the sending machine to send less data next time. This takes care of end-to-end flow-control issues, but shortly we shall discuss another flow-control issue, that of congestion in the network.

The *checksum* detects errors in the TCP packet. The last part of the non-optional TCP header is the *urgent pointer*. It can be used to notify the receiving system that urgent data is coming and that the receiving system needs to act on that data as quickly as it can. The value in the field indicates the end of the urgent data. For example, suppose you are receiving something from a sender and determine that your machine is not reacting properly, so you need to stop the transmission. You send a "stop" command to the server sending the information, and the urgent flag is set to tell the server to act on that packet as quickly as it can.

The *TCP options* field, if used, can carry certain information that may or may not be needed. These options include a window scale factor, time stamps (for time-critical data), and a message digest of the packet. Application data makes up the rest of the data in the TCP packet.

When a host needs to send TCP data, it must first set up a session with the opposite end (recall that a characteristic of connection-oriented protocols is that a session, or *call*, must be set up before data can be passed and must be terminated, or *torn down*, at the end). The process of setting up the session is initiated by the host, which sends a *connection request* message to the intended receiver. The intended receiver then returns a *connection confirm* message, after which the sender returns an *acknowledgment* message. At the completion of this so-called *three-way handshake*, the connection is set up and data may be passed. Similarly, when the information has been passed, the connection must be closed in each direction.[20–22]

Congestion Control

Several times we have referred to congestion and control of congestion. Congestion occurs when something in the network gets data faster than it can process it. The receiving host can become congested if it gets data faster than it can handle it, and this situation is accommodated by the window field described earlier in the TCP section. If, on the other hand, a router gets congested, it can, if possible, set the ECN bits in the IP header, which will cause the receiver to slow down data delivery.

Two common techniques for congestion control at the TCP level are called *slow-start* and *multiplicative decrease*. Notice the transfer rate while downloading a large file, such as the radio frequency interface (RFI) files from *www.cablemodem.com*. You will see that the file transfer rate starts slowly, increases rapidly, possibly backing off, and eventually reaches a more-or-less steady state. You are watching these congestion mitigation techniques at work.

Slow-start avoids swamping the Internet with additional traffic immediately after congestion clears or when a new connection starts. When the file transfer

starts, the sender transfers data at a relatively slow rate, starting by transferring a single segment and waiting until that segment is acknowledged. When that first segment is acknowledged, the sender transmits two more segments. When they are acknowledged, it transmits two more, then four more, and so on. Eventually either the receiver's limit is reached or receipt of acknowledgments slows, indicating congestion somewhere along the route.

If congestion is encountered after the file transfer begins, multiplicative decrease is invoked. The number of segments transmitted is halved, and retransmission delayed, for every acknowledgment *not* received. This forces the sender to rapidly cut back to remove congestion, hopefully before anyone's packets are dropped. After the sender determines that no more packets are being dropped, it enters the slow-start sequence again, allowing the flow rate to increase again. In order to further avoid recreating congestion, as soon as the transmission rate reaches half the rate prior to congestion, the rate of increase is slowed.

5.4 Summary

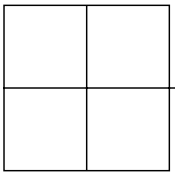
This chapter has attempted to introduce the key points related to data communications on hybrid fiber coaxial networks. It is not intended as a course in data communications, and many important topics in data communications were omitted. Those included were selected because of their particular relevance to data communications as it is being practiced on cable television systems today. One can expect many changes in techniques in the next few years as the technology matures.

The next chapter will deal with one particular system that transmits data on HFC plant, the cable telephone system. Many of the concepts presented in this chapter will be useful there.

Endnotes

1. Larry Peterson et al., *Computer Networks*. San Francisco: Morgan Kaufmann, 1996, p. 35ff.
2. Jim Stratigos, *ISO/OSI Model Introduction and Ethernet Overview*. Preconference tutorial presented at the 1997 Conference on Emerging Technologies, Nashville, TN, by the Society of Cable Telecommunications Engineers.
3. Jean Walrand and Pravin Varaiya, *High-Performance Communications Networks*. San Francisco: Morgan Kaufmann, 1996, p. 92ff.
4. An interesting historical perspective on the Internet, which explains the routing concept somewhat incidentally, may be found in Katie Hafner and Matthew Lyon, *Where Wizards Stay Up Late*. New York: Simon & Shuster, 1996.

5. http://www.cablemodem.com
6. Society of Cable Telecommunication Engineers, *Data-Over-Cable Systems Radio Frequency Interface Specification 1.1*, SCTE 23–1 2002.
7. http://www.usb.org
8. http://www.pcisig.com
9. See http://www.usb.org, *PCI-X 2.0 Overview*.
10. Cisco Systems, *Migrating Simple Data over Cable Services to DOCSIS 1.1*. Available at www.cisco.com.
11. Erich Arnold, *Content on Demand*: *How DOCSIS 1.1 Dynamic Service Flows Deliver the Goods*. Western Show, 2001.
12. Jean Walrand and Pravin Varaiya, op cit.
13. http://www.ieee.org
14. ANSI/TIA/EIA 570-A, *Residential Telecommunications Cabling Standard*. Available at http://global.ihs.com. See Figure 2 for an example of proper residential wiring. See also Endnote 15.
15. BICSI, *Residential Network Cabling*. 2002. BICSI is a not-for-profit telecommunications association supporting the design and distribution of telecommunications wiring in buildings. See www.bicsi.org. The book may also be obtained from commercial booksellers.
16. Charles E. Spurgeon, *Ethernet: The Definitive Guide*. Tampa: O'Reilly & Associates, 2000. Sebastopol, CA.
17. David Cunningham and William Lane, *Gigabit Ethernet Networking*. New York: Macmillan Technical, 1999.
18. Stephen A. Thomas, *IP Switching and Routing Essentials*. New York: Wiley, 2002.
19. http://www.ehsco.com/reading/19960515ncw3.html. See also www.webopedia.com.
20. Stephen A. Thomas, op. cit.
21. Douglas I. Comer, *Internetworking with TCP/IP*, 3rd ed. Upper Saddle River, N.J.: Prentice-Hall, 1995.
22. Floyd Wilder, *A Guide to the TCP/IP Protocol Suite*, 2nd ed. Boston: Artech House, 1998.



Chapter 6
Cable Telephony

6.1 Introduction

This chapter covers the principles of telephony over a hybrid fiber-coax (HFC) network. Principles of telephone operation are discussed for the benefit of cable TV engineers not familiar with the telephone network. The discussion of telephone technology concentrates only on those principles that aid in understanding the operation of telephony on an HFC network. The chapter is not intended to be a dissertation on the operation of the telephone system.

The elements of a modern telephone system, generally from the switch to the subscriber, are outlined in order to aid in understanding the telephone-side interface of an HFC telephony system. This is followed by a brief introduction to the digital hierarchies used in telephone interfaces, which will help in understanding the headend interfaces. A sample HFC telephony system is then described. It is not intended to represent any of the telephone systems on the market but is intended to outline some of the issues of importance to understanding the operation of any HFC telephony system.

Early cable telephone systems were built on switched circuitry technology, and many such systems remain in service. Newer installations tend to be voice-over-Internet protocol systems, described later in the chapter.

6.2 Modern Telephone System Architecture

Today's telephone systems may be broken into several elements. A convenient way to look at the telephone system is to start with the switch, the device that connects telephones. We could look at the side of the switch that interfaces to other switches, or we could look at the side that interfaces to telephones. We spend most of this chapter looking at the telephone-side interfaces since this is where changes are made to accommodate cable telephony. We do, however, mention the other side for the benefit of cable engineers who are not familiar with the telephone system.

6.2.1 The Local Exchange

Modern telephone systems in North America tend to be constructed in a relatively "flat" architecture, as illustrated in Figure 6.1.[1] Previously, a different architecture with a hierarchy of five different types of switches was used, and

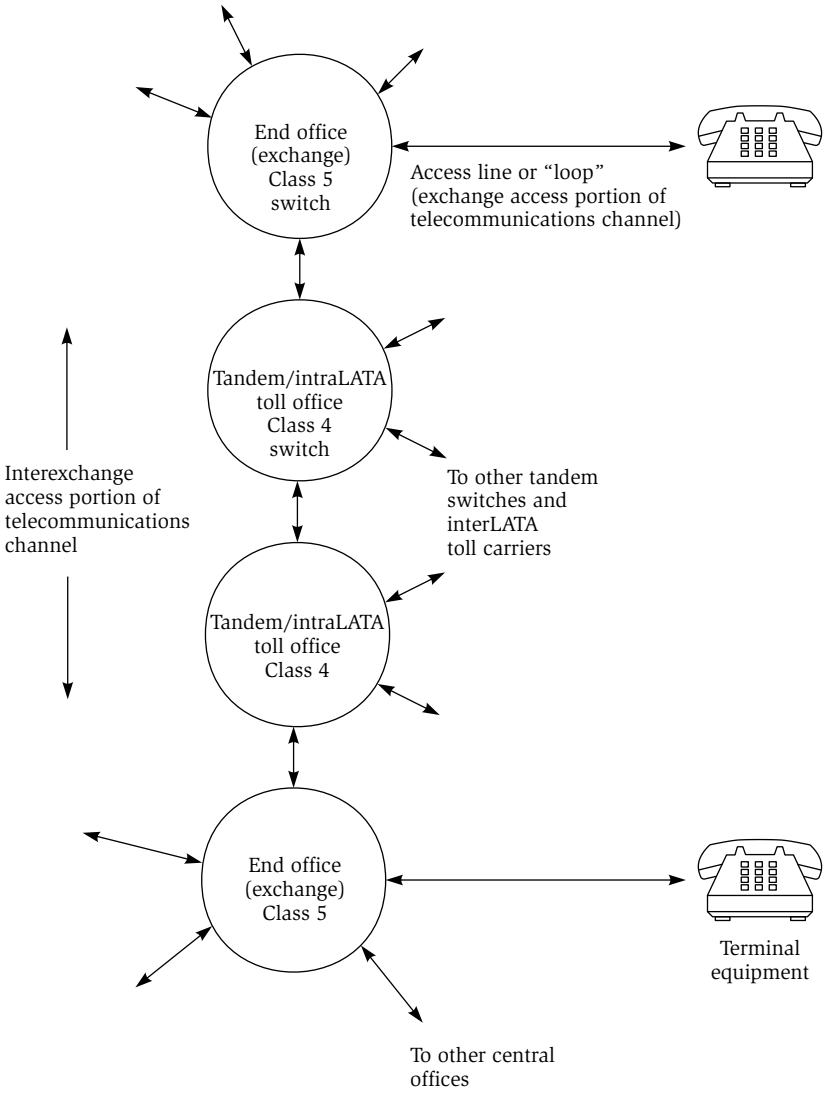


Figure 6.1 Modern telephone system architecture.

some of the terminology is left over from those architectures. In Figure 6.1, each telephone is connected to an *end office*, also called a *central office*, *local exchange*, or *class-5 local exchange*. Local exchanges are identified by the first three digits of the seven-digit local telephone number. If two telephones connected to the same end office talk to each other, then the signal goes no farther than that end office. End offices connect to other end offices and to *tandem offices*, or *intraLATA toll offices*. Tandem offices are the location at which operator services are normally provided ("provisioned" in telephone terminology). These tandem offices connect calls to other end offices and also to long-distance carriers, called *interLATA carriers*. Note the distinction between intraLATA and interLATA.

End offices employ so-called class 5 switches, which are designed to support subscriber features described later in the chapter. Class 4 switches are optimized for long-distance trunking and are used in tandem offices.

The term *LATA* stands for *local access and transport area*. It is the area in which a local telephone company transports long-distance calls itself without passing the call to a long-distance company. If a call is between telephones in the same LATA, then the local telephone operating company carries the call on its own facilities (intraLATA) rather than handing it to a long-distance carrier (interLATA).

In North America, switches communicate with each other using a protocol known as signaling system number 7 (SS7). This is an out-of-band signaling system used to provide basic routing information, call setup, and other call termination functions. It operates on a data network independent of all voice channels.

6.2.2 Switch Architecture

Figure 6.2 illustrates the architecture of a modern digital switch. The set of peripheral interfaces shown at the bottom of the figure convert external signals to digital form if they don't already exist that way. They also perform some low-level call processing. Signals are passed to the *switching matrix*, which performs the function of connecting each call as appropriate. Within the peripheral interface set, *line controllers* interface with individual telephone instruments where they are routed into the exchange. These are ordinarily analog inputs, which must be converted to digital. *DLC interfaces* connect with digital loop carrier systems, a concentration system explained in Section 6.2.3. *Trunk controllers* interface with other end offices and tandem switches.

Voice channels are handled digitally in modern switches. The basic digitization method converts each voice channel in each direction to a 64-kb/s datastream. This is explained in Section 6.3.1.

The input/output controller system provides access to the switch for maintenance, billing, routine operations and administration, and loading of software. The central processor is usually duplicated to ensure reliability. It controls call-processing activities, such as assigning time slots, and administering features, such as call forwarding.

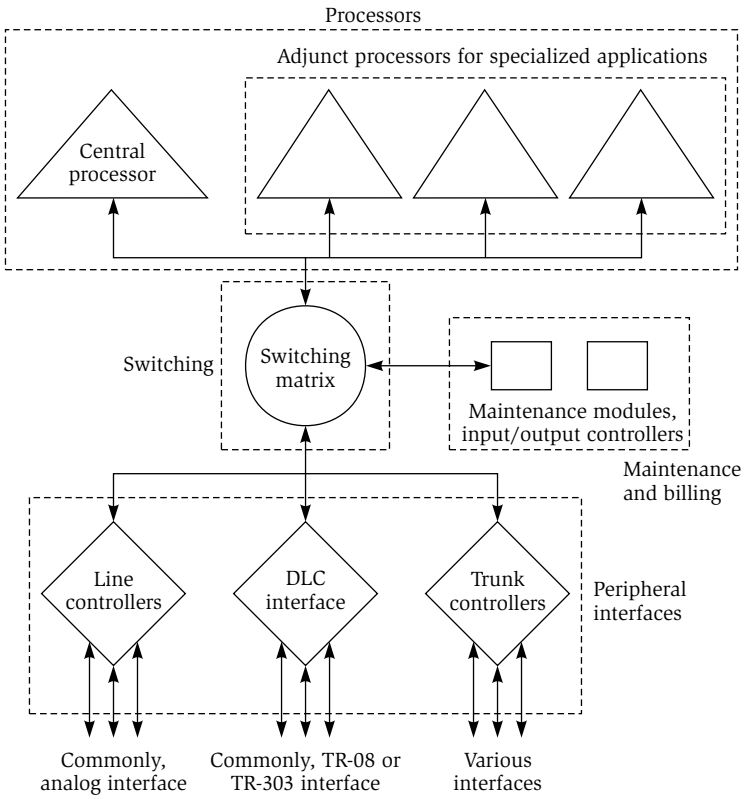


Figure 6.2 Architecture of modern switch.

Call Processing

It is useful to understand what information originates in the switch, and in what form, since this aids in later understanding what functions can be provided in a particular cable telephony system. The switch provides a number of services to the call, as follows:

Call Detection. The switch determines that the telephone receiver has been lifted ("goes off-hook"). This is normally done by detecting current on the line. The current is set up by a voltage supplied from the switch. When the telephone goes off-hook, a resistance is placed across the phone line. The switch's central processor confirms that the phone is a valid subscriber, and marks the line as busy (as opposed to idle). It creates a *call register* to receive and store dialed number data, and a *device register* to handle additional messages to and from the line.

Dial Tone. When a switch detects an off-hook condition, a dial tone generator in the switch provides dial tone to the line.
Digit Collection. The dialed digits are collected, based on either the tones or dial pulses received from the telephone instrument.
Digit Translation. The dialed digits are accumulated until enough digits have been collected to identify where the call is to be routed. If invalid information is detected, the call is directed to a special tone or recorded announcement.
Call Routing. The call is routed to either another phone in the exchange, or another exchange, or a tandem/interLATA toll office, as appropriate. Two time slots in the appropriate interface multiplex are reserved, one for each direction of the call. (A *multiplex* is the sharing of a single communications path — usually a time slot as described in Section 6.3.3 — by multiple calls. As used in this context, it refers to reservation of an appropriate communications path from the originating switch to the next point to which the call must be routed.) If no appropriate time slots are available, the caller gets a tone or announcement advising him or her that a circuit is not available. Often the notification is in the form of a "fast busy" signal.
Call Connection. The terminating circuit (the *called line*) is marked "busy" so that no other callers can terminate on that line.
Audible Ringing and Ringback. The receiving switch signals the called number by placing a ringing signal on the terminating line. A number of different ringing signals have been used in North America, but at this time the most common is 20 hertz, 60 to 108 volts rms. If a DLC or similar system (including cable telephony) is used, the switch commands the line interface at the DLC to generate the ringing signal. A *ringback* signal is sent to the initiating telephone to tell the caller that the phone is ringing.
Speech Path Established. When the called phone is picked up (goes off-hook), the voice paths are established through the previously reserved time slots. Ringing and ringback are removed from the lines. The central processor then causes the connection to be completed.
Call Termination. When a phone hangs up, the loss of current on that line is detected. The call is "torn down" by releasing the circuits, time slots, and so on that were involved in making the call.

Because of system design, it is possible that up to four seconds will elapse between the time a called line is seized and the time a ringing tone is applied to it. This explains why you occasionally pick up a phone to make a call and are surprised to find someone calling in even though the phone didn't ring.

6.2.3 Digital Loop Carrier

Traditionally, telephone instruments have been connected directly to the switch in the local exchange. However, this requires a lot of copper wire to run between homes and the exchange. A more efficient way of interconnecting telephones and switches has been developed in the form of *digital loop carrier*

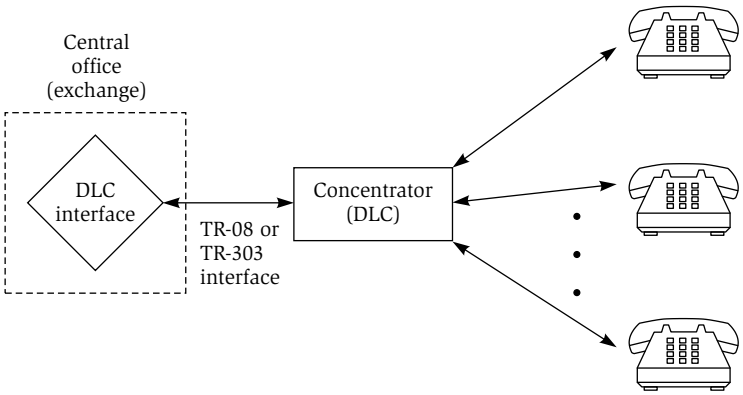


Figure 6.3 Integration of digital loop carrier into system.

(DLC) systems. Bulk transport is used between the exchange and a point in the field closer to the users. From that point, individual copper pairs serve each subscriber.

Figure 6.3 may be compared with Figure 6.1, which showed a telephone connected directly to the switch at the exchange. A DLC system adds a *concentrator* in the field, which multiplexes a number of subscriber lines onto a single output line going to the switch. The concentrator interfaces with the DLC interface at the lower center of Figure 6.2. The DLC converts the analog telephone signals to digital, as do the line controllers in the switch. Rather than being located in the switch, the DLC is located in the field, close to the subscriber.

Two interfaces are commonly used for DLC systems in North America; other interface specifications are used elsewhere. The more established interface is TR-08,* which is a DS1 interface (see Section 6.3.4) consisting of 24 time slots, each of which can handle one telephone call. The newer interface specification is TR-303, a wire line and fiber-optic (SONET) interface. It can support up to 2,048 lines and many telephone features.

Modification of the DLC Concept to Accommodate Cable Switched Circuit Telephony

Figure 6.4 illustrates the way in which a DLC architecture is modified for inclusion of switched circuit telephony over cable television systems. Later

* TR stands for *technical reference*, a Bellcore-created technical document with proposed generic requirements for products, interfaces, technologies, or services.

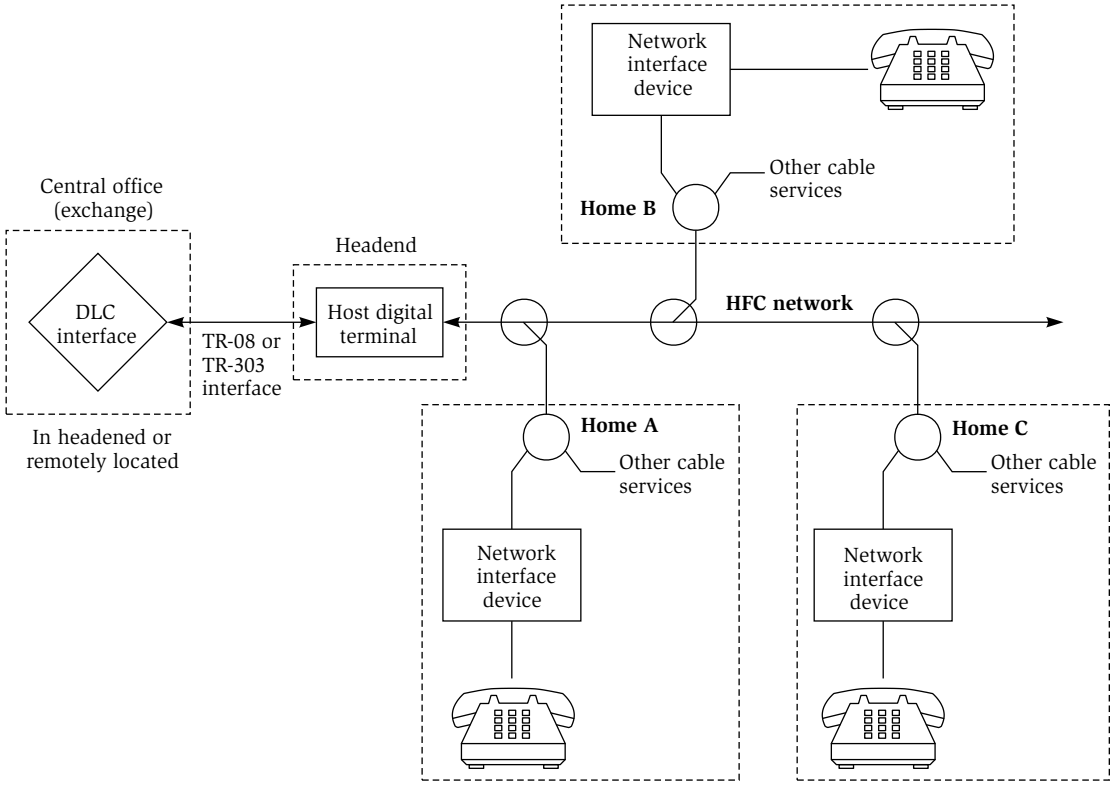


Figure 6.4 Modification of DLC architecture for cable switched circuit telephony.

sections will cover more information relating to cable telephony. Section 5.6 will cover IP telephony.

Compare Figure 6.4 with Figure 6.3. The concentrator of Figure 6.3 is located in the field and collects calls from a number of lines, concentrating them onto one cable for interface with the switch. The host digital terminal (HDT) of Figure 6.4, in the headend, serves the same function and "looks" to the switch as if it were a concentrator. In turn, the HDT interfaces with the telephones connected to the hybrid fiber-coax (HFC) network.

An interface is required at each home subscribing to cable telephony services. This home interface supplies analog signals to the telephones in the home, just as would a switch or a DLC system. It is known by many names, but here we use the generic term *network interface device* (NID) to mean the equipment at the home that interfaces between the HFC network and the subscriber's telephone

instruments. Other names by which the interface is known include network interface unit (NIU), customer access unit, and wallbox.

NID is also sometimes used to mean a plastic or metal box, usually mounted on the side of a house, that contains protection equipment and the subscriber interface port to the telephone system. The term is used in the cable industry at times to describe a similar facility used to house the ground block, maybe a splitter, and maybe an amplifier. It is possible, but not mandatory, to place the telephony interface equipment in a NID according to this definition. In this chapter, we use the term to mean the equipment used to interface between the subscriber telephone equipment and the cable plant.

In general, you would expect to split the cable in front of the NID in order to provide other cable services. The splitter may be provided as part of the NID, or it may be supplied externally. It is possible, and some say desirable, to place a high-pass filter in the "other cable services" side of the splitter output to prevent leakage of spurious signals into the cable plant. This topic is covered in Chapter 16.

6.2.4 Backup Powering

The telephone industry tends to power from a −48-volt dc battery that is float charged from commercial power so that the batteries take over when commercial power fails. Some systems are powered from −130-volt dc. In addition, remote access vehicles are available that have on-board batteries to ensure the survivability of power if commercial power is lost for an extended time. Central offices and major field installations have generator backup and battery backup sufficient to hold the office during generator start-up.

6.2.5 POTS Through CLASS

A number of types of telephone service are available today. For residential service, we can talk about the services that are universally offered to all subscribers, as well as suites of upgraded services. The classical services, or plain old telephone services (POTS), include rotary and tone dialing, voice connection, operator services, access to long distance and international long distance, and in most locations in North America, 911 emergency service. (In most of North America, it is possible to summon police, fire, ambulance, or other emergency services by dialing 911. An emergency operator is presented with a screen showing the location from which the call is placed and pertinent information relating to that location.) We would expect any cable-based telephony system to provide POTS.

Advanced features allow more services and new revenue streams for telephone system operators. A second set of services, custom calling services (CCS), also called custom messaging services (CMS), work with any touch-tone telephone. The services include call waiting, call forwarding, three-way calling, speed dialing, added directory number (often called teen service) with distinctive

Table 6.1 CLASS features

Display Features	
Calling number	Transmitted during ringing
Calling name	Local or SS7 database
Security Features	
Customer-originated trace	Recipient of harassing call dials code and makes caller's number available to law enforcement
Calling number delivery blocking	Allows caller to prevent his or her telephone number from being available to called party
Call Screening Features	
Selective call acceptance	Only calls from certain telephone numbers can ring phone
Selective call rejection	Certain numbers cannot ring phone
Selective call forwarding	Forwards calls only from a list of numbers
Anonymous caller rejection	Rejects calls from anyone who has blocked number delivery
Distinctive ringing/call waiting	Distinctive signals from designated phone numbers
Convenience Features	
Automatic callback	Lets caller know when busy called line is free
Automatic recall	Allows called party to quickly return missed call

ringing, multiline custom calling, and extension-bridged services. Some or all of these features may or may not be supported by cable-based telephony systems.

Custom local area signaling services (CLASS) are a set of advanced services provided by modern digital switches and the expanded intercommunication between them. These services include automatic callback, automatic recall, calling number delivery (also known as caller ID), customer-originated trace, distinctive ringing, selective call forwarding, and selective call rejection. Table 6.1 summarizes CLASS features.

Many CLASS services depend on delivery of the number of the calling party. The number is delivered to the called phone before it is answered. This necessitates transmission of the calling number between applications of the ringing voltage.

Besides POTS, custom calling, and CLASS, other services are often provided by local telephone companies. Network-based voice mail takes the place of answering machines, with the added advantage of being able to answer incoming calls when the phone is in use. Voice messaging, voice menus, fax messaging, and voice forms are offered, primarily to provide small businesses with big-business types of features.

Integrated services digital network (ISDN) services allow digital transmission on the normal copper pair serving businesses and residential users. So far,

ISDN has had most of its impact on businesses. It allows simultaneous voice and data transmission on the same copper pair. *Basic rate ISDN* operates at 144 kb/s and allows two 64-kb/s voice channels (b, or bearer channels) and one 16-kb/s data channel (d channel). *Primary rate ISDN* operates at nearly 1.5 Mb/s and offers 23 b channels and 1 d channel.

6.2.6 The Subscriber Loop

Here we present more information on the operation of the subscriber loop, the portion of the telephone system that connects to the telephone instrument.[2] This is of interest to the HFC telephony technologist because it describes the interface between the NID and the telephone instruments in the home.

Residential telephone service is provided using a single twisted pair line, which carries signaling as well as voice in both directions. The two wires are referred to as *tip* and *ring* — from the Model 310 phone jacks used to make circuit connections in the days when the telephone system was operated manually. The 310 jack is also known as a *phone jack* and is used for many audio applications. If a third connection is added for stereo, that connection is called the *sleeve*.

Battery

A station battery voltage of 24–48 volts dc is supplied from the central office (the term *battery* is used regardless of the source of the voltage). The *ring* wire is negative with respect to the *tip*, but modern telephone instruments are insensitive to the polarity of the applied voltage. The battery voltage is used to power the telephone instrument and to provide current that indicates an off-hook condition. Current is limited to prevent wire heating and personal injury, and to protect the central office against shorts on a line. In the case of a cable telephony system, the network interface device provides this voltage.

Ringing

When the telephone rings, a sinusoidal alternating voltage is placed on the line, sometimes superimposed on the 48-volt battery voltage. The voltage used varies from about 60–108 volts rms, and would likely be toward the low-voltage end of the range for cable telephone interfaces. The higher voltages are not needed because the telephone does not have to be rung from a long distance, so the voltage drop in the twisted pair is low.

A number of different frequencies have been used for the ringing voltage. For party-line systems, in which multiple homes share one telephone line, one method used to ring the telephone of only the called party is to assign each home a different ring frequency, then to place a filter in each drop, which would pass ringing voltage of only the frequency intended for that home. In North America, party lines are rarely used anymore, so a multitude of ring frequencies is not as important.

A common ring frequency today is 20 Hz. For HFC telephony systems, the ring signal is generated at the NID rather than being transmitted from the headend.

Seizing the Line

When a receiver is lifted off the hook (goes off-hook), a resistance of around 200 ohms is placed between the tip and ring. That resistance and the battery voltage set up a current, which is detected as an off-hook condition by the switch. In the case of a cable telephony system, the off-hook condition is detected by the network interface device at the home. The NID then signals the switch of the off-hook condition.

This system, in which the resistance is applied between the tip and ring, is called a *loop start* system and is by far the most common method of seizing a line in residential telephone systems. Another method, used primarily in certain private branch exchanges (PBX) is *ground start*, in which the resistance is placed from either the tip or ring to ground. We are not likely to encounter such a system in cable telephony work.

Dialing

Two methods are used to communicate the called number to the switch. The older method uses a rotary dial. Every time a digit in the telephone number is dialed, the dial is pulled to a stop and released. As it returns to its rest position, it breaks then remakes the circuit a number of times equal to the number being dialed. Recall that as soon as the telephone goes off-hook, a load is placed on the loop, that is, the circuit is "made." The circuit is "broken" by momentarily removing the load, returning the circuit to its on-hook condition. The nominal duty cycle of one break/make is 0.1 second. The circuit is broken for 60 ms and made for 40 ms. A period of 200 ms minimum represents the spacing between digits.

The method of "dialing" used more commonly today is the use of dual-tone multifrequency (DTMF) signaling. These are the well-known touch tones. For each digit, 0 through 9, two sinusoidal tones are generated simultaneously. One tone is from the so-called low group (representing rows on the telephone keypad), and the other is from the high group (representing columns). A total of 16 states can be represented by the tones, of which 12 are normally available on consumer telephone instruments. The last four tones, reserved for special functions, are not used for customer premises equipment.

At the switch, tones are detected using a tone decoder, which supplies digit information to the call register. As the number is dialed, it is interpreted by the switch based on the information placed in the call register. In an HFC telephony system, the tones will be passed through the NID and to the switch.

Two tones are used to communicate each number. Though it is possible to communicate the same information using single tones (16 frequencies would be needed), use of dual tones improves reliability. A single-tone detector could be more easily confused into thinking that a touch tone was being transmitted when

it was not. It is much less likely that a transmitted signal would simultaneously include power at the two required frequencies.

Call Progress Signaling Tones

As the call progresses, a number of tones are used to effect signaling and to indicate the status of the call. Figure 6.5 illustrates the call progress tones commonly used in North America, and also includes the DTMF tones. Table 6.2 describes each call progress tone, as well as a few other tones of interest.

Sidetone and Echo

Though the telephone transmission system is a two-wire system with signals going both ways on the same wire pair, within a telephone instrument, the signals are separated, using a hybrid transformer or an electronic equivalent. The hybrid by itself can provide excellent isolation between the transmit and receive directions. In fact, the isolation can be too good: a talker who doesn't hear his or her voice at the correct level in the receiver will tend to shout. For this reason, it is common in designing telephone instruments to intentionally couple some energy from the transmitter (mouthpiece) to the receiver (earpiece). This is called *sidetone*.

Sidetone is desirable, but it is similar to a phenomenon that is undesirable: echo. Echo occurs when a signal travels from the sending telephone to the

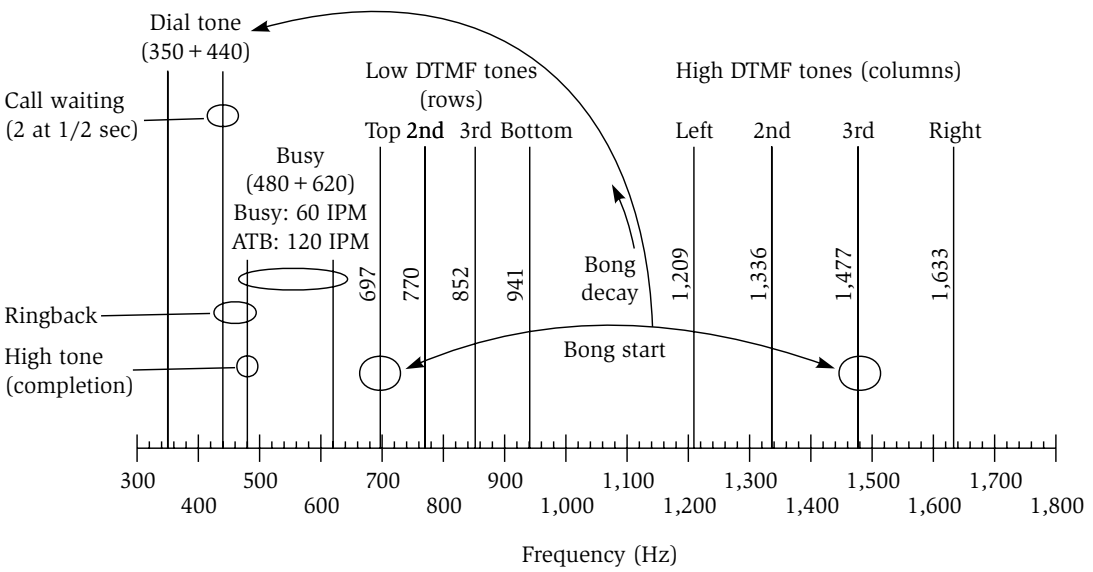


Figure 6.5 Call progress and DTMF tones.

Table 6.2 Call progress tones

Signal	*Tones (North American plan)*	*Description*	*Alternative Plans*
Dial tone	350 and 440 Hz	Placed online when switch detects off-hook condition	600 and 120 Hz, or 600 and 133 Hz
Busy signal	480 and 620 Hz interrupted 60 times per minute (60 IPM)	Signifies calling line is in use	Dial tone frequencies interrupted 60 IPM
All trunks busy (ATB)	Same as busy except 120 IPM	Switch is unable to complete the call because of unavailability of resources	
Ring-back	440 and 480 Hz, 2 s on, 4 s off	Informs caller that called phone is ringing	420 and 40 Hz
High tone	480 Hz	Confirmation (e.g., phone programming option has been completed successfully)	500 Hz
Bong	697 and 1,477 Hz followed by exponentially decaying dial tone	Prompts for entry of calling card number	
Off-hook	1,400, 2,060, 2,450, and 2,600 Hz	Tells subscriber that phone is off-hook	

receiving telephone, then is reflected to the sending telephone. The caller hears his or her own voice delayed. In long circuits, it is common to use echo suppression to reduce the intensity of the echo. Echo suppression may be very simple, such as increased attenuation in the direction not speaking, or more complex signal-processing circuits can be used. In the latter case, the term *echo cancellation* is often used.

In shorter circuits, echo suppression is not used. In order to prevent these circuits from creating objectionable echo, Telcordia has established a standard for delay in the circuit.[3] The Bellcore standard for round-trip delay of a voice signal is given by

$$RTD_{\max} = 0.32 + 0.0168M$$

where

RTD = round-trip delay (in milliseconds)

M = one-way route distance (in miles)

This requirement may be broken into two parts: the 0.32 ms allows processing delay at each end of the circuit. This limited processing delay is the reason some manufacturers of cable telephony equipment are reluctant to use error correction in the cable transmission: error correction could put them outside the delay requirement because of the way that practical error-correction systems work. Error correction is applied over a certain number of bits. The bits must be held at the receiver until all are received; further, they must be examined for errors, and any errors must be corrected. Then those bits can be released for the next processing step. Moreover, most practical error-correction schemes depend on the order of the bits being changed so that a burst of error will not affect a packet beyond the point where errors could be corrected. This rearrangement of the order of the bits adds more processing delay.

The second element in the equation is the propagation delay in the medium, be it fiber or coax. The round-trip delay permitted by the equation is 16.8 μs per route mile. Since a delayed signal traverses the path twice, one can allow propagation delay of 8.4 μs per mile of plant. This is barely adequate for modern fiber-optic cables, and indeed may be slightly below the propagation delay in some cables today. Of course, propagation in coaxial cable is faster, so the coax portion will exhibit less delay.

Line Balance

Telephone signals are normally transmitted on twisted pair lines using balanced transmission. This is contrasted to the unbalanced transmission of most RF signals. In order to control noise pickup on telephone lines, it is important to control the balance between the two wires. We can show that if a transmission path is truly balanced, it neither radiates nor picks up power from the outside world. This follows from Ampere's law, which states that the magnetic field surrounding a conductor is proportional to the *net* current in that conductor. If we observe from such a location that two conductors having equal but opposite currents appear to be in the same place, then no magnetic field exists due to the current. Real balanced paths are, of course, a compromise in that it is not possible to totally balance two wires of a transmission path. However, the better balanced a path is, the less it will radiate or pick up signals.

In order to improve the balance of a circuit, it is useful to twist the two wires making up the circuit. This forces them to be in proximity to each other, which limits any magnetic field radiation (or pickup). Because the wires are often twisted, telephone engineers often refer to a wire pair connecting to a telephone as a twisted pair. The same issues apply to the handling of audio signals in headends, and is covered in more detail in Chapter 8.

6.3 The Telephone Network Digital Hierarchy

We now discuss the telephone network on the transmission side of the switch. This is the side of the switch that connects to other switches, not to telephone

sets. The telephone network has been extensively digitized since the 1960s, based on a standardized hierarchy. Digital transmission was introduced in the 1960s, followed by digital switching in the 1970s. The starting point is the digitization of a single telephone call. The digital resources required for a call are identified, then the multiplexing of a number of such calls is considered.

6.3.1 Sampling of the Voice Signal

The first step in digitizing a voice signal is to sample it. The samples are taken in order to give an analog-to-digital converter a fixed level to sample. In order to promote good-quality voice transmission, it is known that you must transmit, as a minimum, frequency components from about 300 to 3,000 Hz. A somewhat wider range adds naturalness to the voice. It was shown by Harry Nyquist of Bell Laboratories that the minimum number of samples per second that could be taken and still allow the signal to be reproduced was equal to twice the highest-frequency component of the signal.[4] It is essential to filter out any components exceeding one-half the sampling rate. If a higher-frequency component is sampled, then the spectrum of the sampled signal will fold back on itself, resulting in "aliasing" and a very distorted recovered signal.

The signal is sampled at some instant of time, and the sampled level is held in the capacitor, shown in Figure 6.6, for the time required to digitize the voltage level. Then a new sample is taken, and the process begins again. At the top, the signal is sampled, digitized, and sent through the "system" (in this case, the telephone system), then converted back to analog. The sampling process may be accurately modeled as an amplitude modulation process, in which the "carrier" is a square wave that goes between the value 0 (switch open) and 1 (switch closed). The rate at which the signal is sampled in the telephone system is 8,000 samples per second (8 ks/s), resulting in a Nyquist bandwidth, the bandwidth that can be reproduced, of 4 kHz. The practical upper frequency of the voice signal is lower than 4 kHz so as to allow for filter rolloff.

The spectrum generated by the sampling process is shown at the bottom of Figure 6.6. The baseband spectrum is present even after sampling. In addition, the first harmonic of the "carrier," at 8 kHz, is shown, surrounded by the normal modulation sidebands, which are due to the audio. The second harmonic, at 16 kHz, is also present with sidebands, though not shown here, and so on. Note that the spectra of the baseband signal and the harmonics do not overlap.

6.3.2 Analog-to-Digital Conversion

After sampling, the voice signal is converted to a digital representation. It has been found that a suitable minimum signal-to-noise ratio for voice transmission is approximately 30 dB. (The definition of signal-to-noise ratio is somewhat different from that used in video systems.) We can analyze the signal-to-noise ratio of a digitization process by considering the effect of an error in the least significant bit (LSB) of the digitized signal. It may be shown that a system for digitizing analog

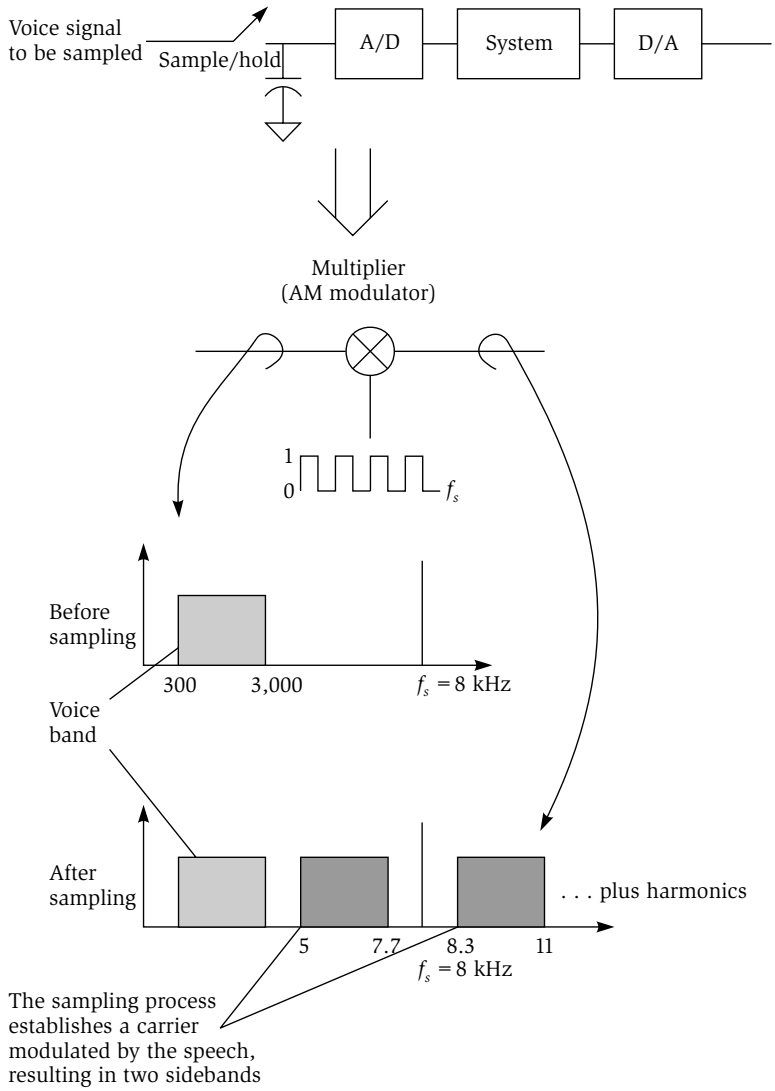


Figure 6.6 The sampling process.

signals exhibits an encoding error ("quantizing error") of one least significant bit. This is taken as the quantizing noise of the system. Quantizing noise acts somewhat the same as the random thermal noise familiar to cable engineers in that it masks the true value of the signal by creating an uncertainty in the value.

Telephone signals are digitized to eight bits. Since $2^8 = 256$, this process represents the analog signal as 1 of 256 levels. One minimum increment is 1 of these 256 levels, so the quantizing signal-to-noise ratio is

$$\frac{S}{N} = 20 \log \frac{1}{256} = -48 \text{ dB}$$

This provides an acceptable safety margin from the minimum deemed necessary. In order to further improve voice quality, the analog-to-digital conversion process utilizes so-called *mu-law encoding*, in which the digital states don't represent constant voltage increments of the analog waveform. Low-level signals are encoded with less distance between states than are high-amplitude signals. By doing this, we can gain apparent signal-to-noise ratio with fewer bits. Various encoding laws (values of mu) are used in different countries. Outside North America, the A-law is used for encoding.

6.3.3 Time Division Multiplexing in Telephony Systems

As we stated, a voice channel is sampled 8,000 times a second, and each sample is encoded to 8 bits. The data rate required to transmit the signal is thus $8 \times 8{,}000 = 64$ kb/s (kilobits per second). This data rate is a basic rate of the time division multiplexing hierarchy used in digital telephony systems. The rate is called the *DS0* (pronounced "D S zero"), where DS stands for digital signal. Thus, when a telephone engineer speaks of a DS0, he or she means a 64-kb/s data signal.

A number of signals may be multiplexed to allow carriage of many signals on a single transmission path. In the standard North American hierarchy, 24 DS0s are multiplexed together to form a DS1, which operates at 1.544 Mb/s. The 24 DS0s require 1.536 Mb/s (megabits per second), whereas the remaining 8 kbps are overhead. This DS1 datastream is often called a T1 datastream, though technically T1 applies only when the transport medium is a twisted pair of wires.

European telephone companies tend to use a different multiplexing strategy in which 30 calls are multiplexed in a 2.048-Mb/s datastream, known as an *E1* stream. The data rate of 2.048 Mb/s is fast enough to allow two more voice channels, but those channels are reserved for signaling (network control). The E1 data rate is sometimes used in North America internal to a system.

6.3.4 Asynchronous Digital Hierarchy

Figure 6.7 illustrates the hierarchy used in North America for digital multiplexing of asynchronous digital signals.[5] The basic DS0 datastream may represent one data source, such as a telephone call, in which case it is called DS0-A.[6] It may also comprise several lower-data-rate channels, in which case it is called DS0-B. (The terms *DS0-A* and *DS0-B* are used only for digital data service

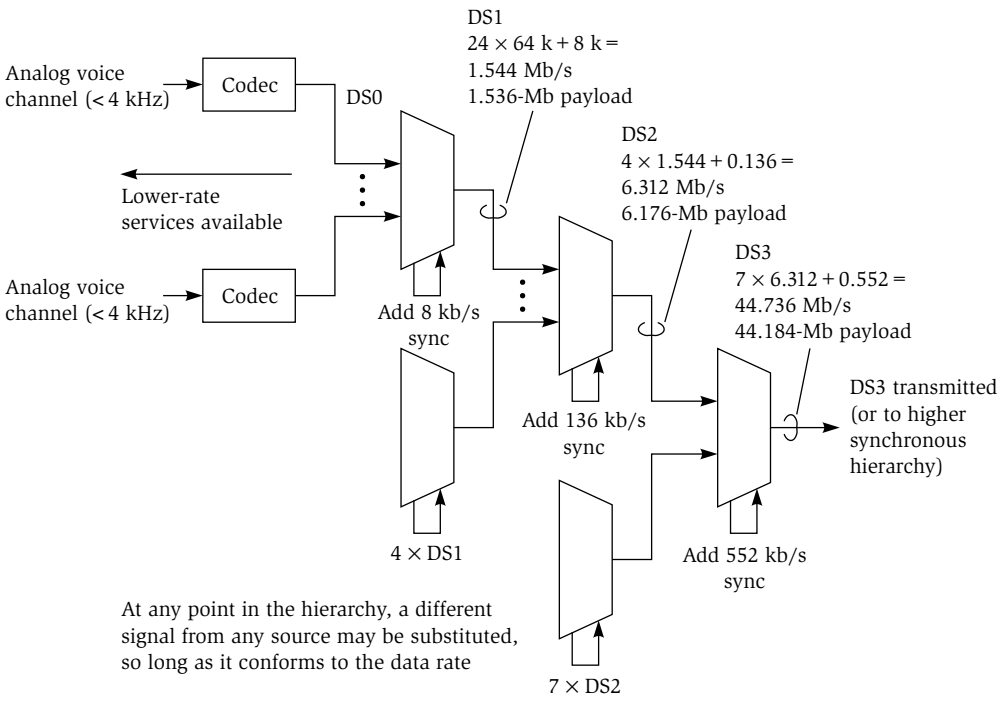


Figure 6.7 North American asynchronous digital hierarchy.

protocols, not for voice work. For voice work, we usually refer to simply DS0.) Twenty-four such signals are multiplexed, along with 8 kb/s of synchronization information, to form a DS1 signal. This 1.544-kb/s signal is very commonly encountered, but is slowly yielding to higher-rate asynchronous signals, and even to higher-rate synchronous signals. Issues related to handling DS1 signals are covered in Section 6.8. Four DS1s may be multiplexed to form a DS2. Only rarely will you encounter a DS2: it is usually just an intermediate multiplexing level in a system that ultimately multiplexes to an even higher level. Seven DS2s may be multiplexed to form a DS3, which can carry 672 simultaneous telephone conversations, or other digital data at the same rate. For example, television networks sometimes use a DS3 to transport a single digital video signal that has been compressed using algorithms most often based on linear predictive coding.

Notice that we have defined only a signal format. We have said nothing about the medium on which the signal is carried. The medium may be twisted pair (at least up through DS1, which is known as T-1 when carried on two

twisted pairs — one in each direction), coaxial cable, microwave radio, optical fiber, or any other suitable medium.

The TR-08 interface to a telephone switch is based on DS1, several of which may be interfaced from one DLC multiplexer. In Europe and other places, a similar interface is known as the channel associated signaling (CAS) interface.

6.3.5 North American Synchronous Digital Hierarchy

A newer format used for optical signaling is known as synchronous optical network (SONET) in North America. In much of the rest of the world, the format varies slightly and is called synchronous digital hierarchy (SDH). The word *synchronous* in this context refers to transmission of signals that have a constant bit rate and are derived from a common clock. It may be that the signals are initially developed from different clocks (though this is unlikely), but then they would have to be synchronized before being multiplexed. It builds on the asynchronous hierarchy shown earlier but is extended to much higher speeds. Extra bits are included to allow the data to "float" in the time slot. This means that if the data arrives after the allocated time slot begins, it can be accommodated if it is not too late, in which case it would wait for the following allocated time slot.

Figure 6.8 illustrates the synchronous digital hierarchy. Although there are variations in the multiplexing options, it is convenient to think of the starting point as a DS3 signal at 44.736 Mb/s. This signal is up-converted to a 51.84-Mb/s datastream by adding some low-rate data and/or "stuffing" extra bits around the payload (the DS3), to add up to the desired number of bits per second. This datastream, 51.84 Mb/s, is known as STS-1 data, where STS stands for "synchronous transport signal." When converted to optical form, the signal is known as an OC-1 signal, where OC stands for "optical carrier." As opposed to cable television practice, the baseband STS hyphen; signal on/off modulates the laser without use of an RF carrier.

Signals can be multiplexed further. If the signals to be multiplexed are in optical form, the multiplexer usually converts them to electrical form to multiplex them, then convert back to optical form. Optical switching and multiplexing is occasionally used. The next data rate is three times the OC-1 data rate, 155.52 Mb/s, and is known as an OC-3. One OC-3 is capable of carrying up to 2,016 telephone calls or an equivalent amount of other data. It is possible to use even higher data rates. Commercial equipment is available at this writing to support OC-192 (9.953 Gb/s) transmission.

The TR-303 switch interface mentioned earlier is capable of operating with either synchronous or asynchronous interfaces. In Europe and some other places, the equivalent interface is known as the V5.2 interface. Unfortunately, V5.2 and TR-303 are incompatible.

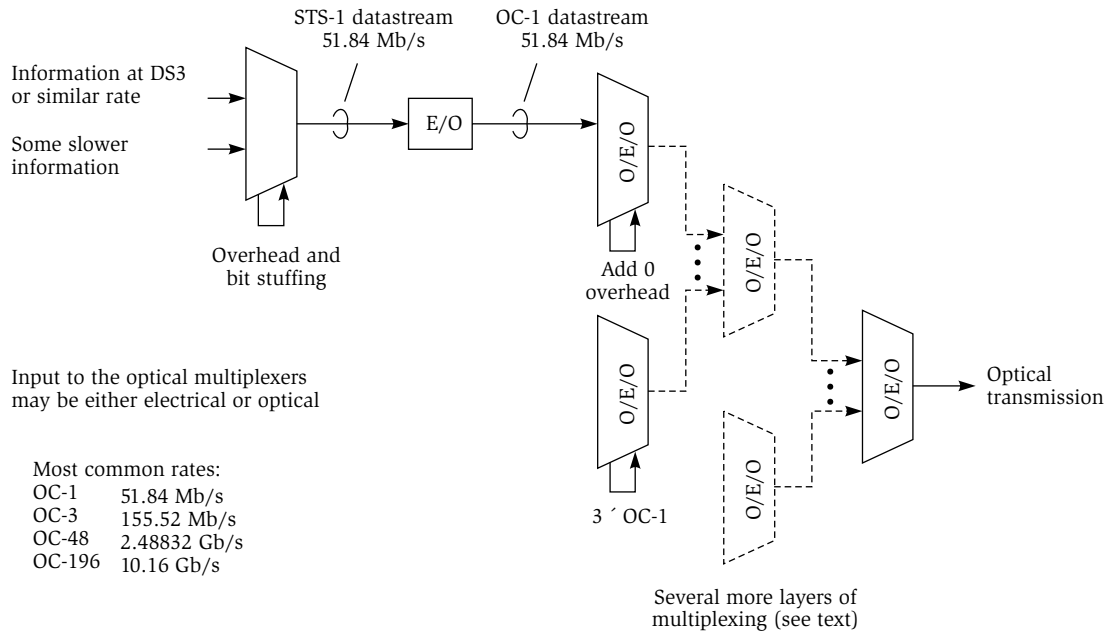


Figure 6.8 Synchronous digital hierarchy.

6.4 Elements of a Cable Telephony System

Figure 6.4 illustrates the overall architecture of a cable switched circuit telephone system. The interface hardware in the headend is generically known as a host digital terminal (HDT). It consists of modulators and demodulators on the cable interface side, and either a TR-08 or TR-303 interface on the telephone side (outside North America, V5.2 interfaces are becoming common with newer switches, and CAS interfaces are common on older switches). Between the cable and switch sides of the HDT are a host of functions required to format signals in both directions, control the HFC link, translate protocols, and support a host of telephone services. The intervening equipment also provides operational support (a common term in the telephone industry for a superset of what the cable industry calls *status monitoring*).

6.4.1 The Home Interface

A transition is necessary between the signals on the HFC network and those required to work with standard analog telephone instruments. This equipment

may be located in the home, but is often mounted on the side of the house, where it forms an interface between the cable plant and all cable services located in the home. The box on the side of the home is called a network interface device (NID). Other names have also been given to the box.

Figure 6.9 illustrates the basic functions of an HFC telephone NID. Also shown is a portion of the HFC network, from the optical node, through a bidirectional RF amplifier, the tap at the subscriber's home, and the NID. Cable signals are routed through the NID, where a directional coupler extracts some energy for the telephony system. The rest of the power in the signal continues to the home to provide other cable services. In this path, the manufacturer may choose to provide a service disconnect switch, an amplifier, and/or a filter to remove upstream energy coming out of the home.

From the directional coupler, the signal is routed to a diplex filter, which separates the downstream frequencies above 54 MHz, from the upstream frequencies, usually below 40 MHz. The high-frequency output from the diplexer is supplied to a receiver, which receives the downstream telephony signal. The automatic gain control shown is conventional and corrects for any variations in signal level as a result of system operating point.

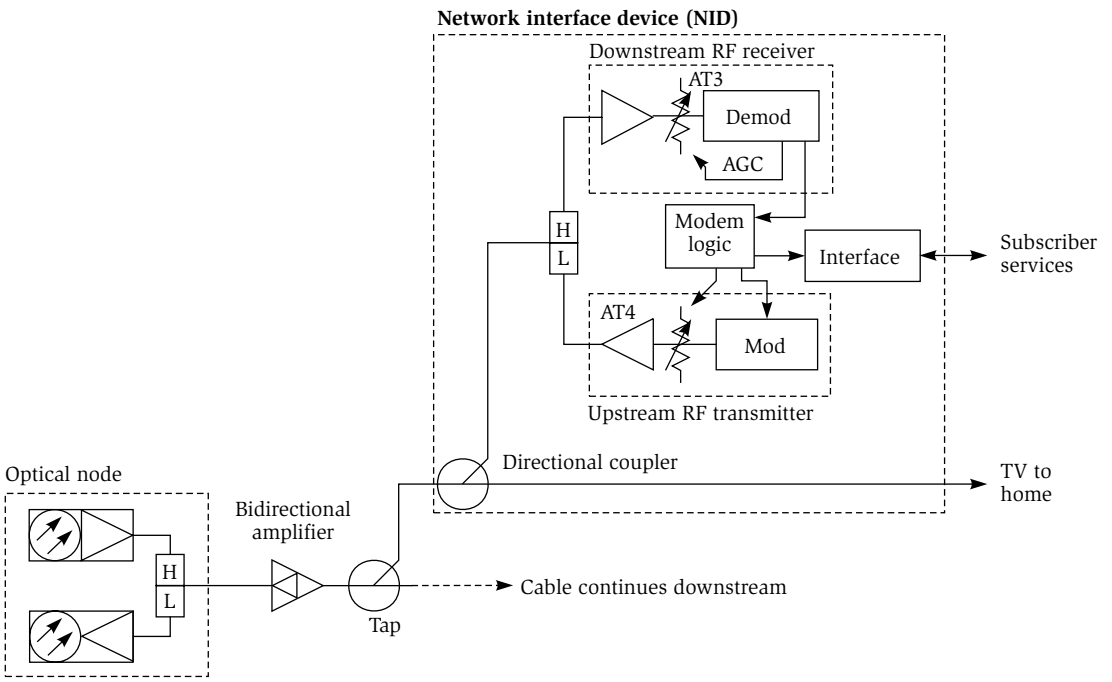


Figure 6.9 Telephone interface at home.

The low-frequency side of the diplexer is used as an input for the upstream signal that comes from the NID transmitter. The attenuator, AT4, controls the output level from the transmitter. As explained in Chapter 16, a long loop ALC system is used, in which the level received at the headend is compared with the expected level, and attenuator AT4 is commanded to adjust the transmitter output level until the level received at the headend equals that expected.

The modem logic is responsible for tuning the transmitter and receiver to the correct frequencies, using time slots (or equivalent) assigned from the HDT, adjusting transmit level according to instructions received from the headend, format conversion, and control of the NID. The interface circuitry provides standard analog signals to the telephone instruments in the home. It provides 24- or 48-volt dc battery voltage, ringing voltage, and dial tone, as well as handling conversion of voice signals between the analog and digital domains.

One or two lines are typically supported in a NID. It is possible in some systems to replace the analog telephone function for one line with a 64-kb/s data channel. A different subscriber interface is required, as is a different interface at the switch.

6.4.2 Powering the NID

A very important consideration involves powering the NID. At the time of this writing, four powering philosophies are being considered by the industry:

1. Powering from the plant using the center conductor of the drop cable
2. Powering from the plant using a twisted pair bonded to the drop cable
3. Powering from the home ac supply, with a backup battery
4. Powering from the home without a battery, and foregoing lifeline service

A principle embedded in the first three powering philosophies is that telephone service should be maintained in the event of commercial power failure, as is done by the incumbent phone company. This requires some sort of battery power and possibly generator power also. The fourth philosophy implies that backup power is too hard and too expensive to supply, so the cable company will not attempt to provide lifeline service and will concentrate on serving the second-line market.

Early generations of NIDs drew 5–10 watts of power. The industry would like to see power drain reduced to 1 or 2 watts. As the power drain decreases, network powering of some sort becomes more palatable. Typically, a NID will draw less power at idle, more while in use, and the greatest power when ringing.

NID Powering Using the Center Conductor

This is a popular method of system powering, where permitted by local code. It can use the existing drop if that drop is in good condition. If the drop is not in good

condition, it should be replaced when adding two-way services anyway. Power is delivered over the center conductor of the drop cable from a special power-passing tap (sometimes called a telephony tap). Chapter 10 discusses power-passing taps.

Of concern with using the center conductor to pass power to the NID is that if arcing occurs, due to a faulty center conductor contact, the arc will transfer a very significant amount of power to the upstream plant, likely causing interference with all users of the reverse spectrum. On the other hand, an arc often causes healing of a bad contact and thus can be self-extinguishing.

NID Powering Using a Twisted Pair

A similar power-passing tap places power on a separate pair of terminals, ground and hot, for each drop. A special drop cable is used that has a pair of wires molded into the same outer jacket that covers the coaxial cable. This cable is often called *Siamese cable*. The tap used with Siamese cable has connectors that are used to fasten the wire pair of the Siamese cable.

The advantage of the twisted pair (which may not actually be twisted) over center conductor powering is that you do not have to tolerate the signal loss associated with connecting RF blocking inductors to the tap ports. The problem with using this configuration is that a special drop cable must be run. In all-new installations this is not a problem, but in retrofit situations, it may not be economical or desirable.

A further concern with Siamese power-passing taps is the possibility of ingress due to signal pickup on the wire pair. The signals picked up can be introduced into the coaxial cable. Adequate filtering of the voltage connections usually is sufficient to prevent problems.

Home Powering with Battery Backup

This method of powering relieves the cable operator of the burden of powering the NID, but requires the use of a battery in each home. This raises issues of long-term maintenance: who is responsible for maintaining the battery and ultimately changing it? What happens if it is the subscriber's responsibility, but he or she doesn't do it and loses an emergency call as a result? In addition, the installer will have to go inside the home and locate a power outlet from which he or she knows the power supply will never be removed. In some cases, this will necessitate installing a new power outlet, a job for a licensed electrician in most locations.

Closely related to the issue of battery maintenance and change-out is the need to know the condition of the battery. Though average battery life is moderately well known, given a battery type and conditions of use, some batteries will inevitably fail earlier, either from unusual environmental conditions or from random failures. By monitoring the voltage of individual battery cells, it is possible to deduce the condition of the battery with acceptable accuracy. Each cell is monitored, and the results will probably be telemetered to a central location.

6.4.3 Headend Telephony Equipment

Figure 6.10 illustrates a headend configuration to support a switched circuit HFC-based telephony system. Variations abound. The downstream HFC plant receives signals from a transmitter, typically part of a bank of transmitters and receivers (modems) dedicated to the telephony system. Typically, the downstream signal will be located toward the higher end of the downstream spectrum, perhaps between 350 and 750 or 870 MHz.

A cable operator may own and operate his or her own switch, or may lease access to the incumbent telephone operator's switch. The transmitter and receiver interface to a telephone switch using one of the interface protocols (TR-08 or TR-303 in North America) described in this chapter. The switch may be co-located with the HFC telephony equipment, or the two may be interconnected with a suitable link, such as a DS1 (TR-08) or SONET (TR-303) interconnection.

Monitoring the Return Spectrum

It is useful to monitor the return spectrum independently of the signals being received by the headend receivers. Return spectrum tends to consist of frequencies that are useful and frequencies that are less useful because of interfering signals. Whatever the source of interference, it is usually the case that some frequencies are not usable at certain times of the day. (Upstream issues are treated in Chapter 16.)

Because of the return spectrum that cannot be used at certain times, many cable telephony systems employ frequency hopping. If a frequency on which the return path is operating becomes corrupted to the point of affecting transmissions,

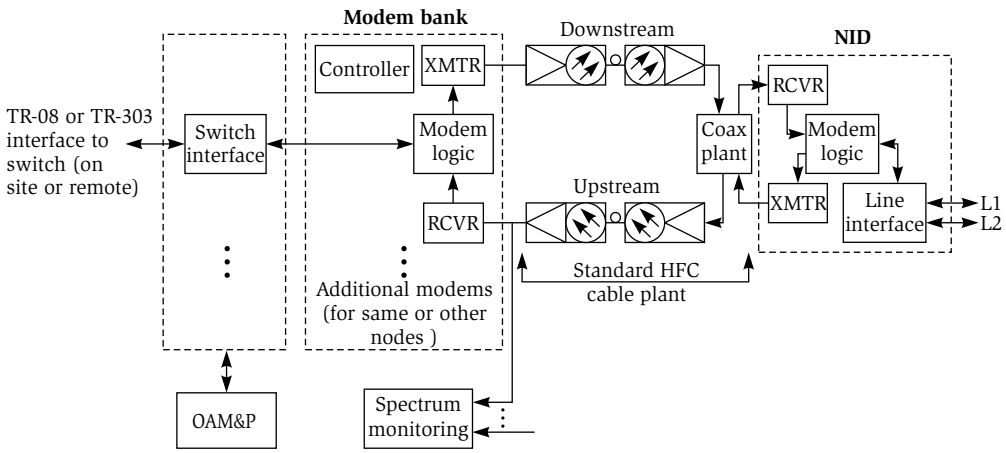


Figure 6.10 Headend elements of telephony system.

the system will switch to a new frequency. If the system does not have the ability to monitor unused spectra, then when it is time to change frequencies, the hop would be made blind. It is possible that the system would hop to a new frequency that was as bad as or worse than the one it left. One way to prevent this is to use a separate spectrum-monitoring receiver, as shown in Figure 6.10.

Typically, the spectrum-monitoring receiver monitors a number of nodes sequentially, allowing the modem bank to have a continuously updated database of quiet frequencies, which are candidates for hopping should the need arise. A threshold is established based on received level at the monitoring receiver. When power is detected above that threshold, the frequency is marked as bad. It is taken out of the database of candidate hop frequencies.

The receiver may serve other needs as well. As shown later in the discussion of the marshaling process for TDMA-based return systems, a time offset must be determined for each NID, based on measurement of the round-trip propagation delay between the headend and that NID. One way to obtain the measurement is to have the NID transmit in response to a command from the headend in a designated time slot on the operational upstream channel. This presents a problem, though, because the delay in fiber-optic cable can easily cause the transmission to be received out of the designated time slot, causing interference with another NID. The monitoring receiver can be used to allow an out-of-channel transmission. From this data, the approximate in-channel time delay is established, ensuring that when the NID is moved to the operational channel, it will not transmit out of its assigned time slot.

Finally, the monitoring receiver can be used to monitor the return paths for any problems that would require maintenance attention. With appropriate software, a flag could be generated if an unusual pattern or level of interference developed. Maintenance personnel would be dispatched to uncover the cause.

OAM&P

OAM&P is a term used in the telephone industry for operations, administration, maintenance, and provisioning. It is a generic term for software suites that allow centralized administering of a telephone network. It is a superset of what the cable industry knows as status monitoring. In the telephone industry, cable's "status monitoring" would be a part of the "maintenance" part of OAM&P. This interface allows for determining the status of the telephone system, downloading software upgrades for any part of the system, and configuring the hardware to match services ordered by subscribers.

6.4.4 Channel Sharing

For the sake of efficiency, it is often desirable to share one communications channel between a number of users. This may be done in many ways, which we

describe in more detail in Chapter 4. The review here is oriented toward a time-sharing protocol, which is common and related to that used in baseband data communications, as outlined in Section 6.3. In the downstream direction, time division multiplexing (TDM) is used. Each of the multiple (typically 24 or 30) lines associated with the data carrier is assigned one time slot. Data is transmitted in time slot 1 to line 1, then to line 2 in time slot 2, and so on.

In the return direction, a related concept is used, except that the multiplexing must be done by turning off one transmitter, then turning on another. This is called time division multiple access (TDMA).

In typical telephony practice, the data in either direction is divided into *packets*, or data sets that are transmitted at one time. Each packet will contain several samples (frequently eight) of a single phone call. The transmission of an entire set of packets from each allocated user is called a *frame* (analogous to the television usage, where a frame is one complete picture). Commonly, frames are repeated every 1 millisecond, a frame rate of 1,000 per second. In a frame, there will be 24 packets (if the system accommodates 24 lines), each packet containing 8 samples of 8 bits each, plus overhead.

6.4.5 A Sample HFC Telephony Marshaling Protocol

This section deals, in a fairly general manner, with some of the issues that are addressed when a new NID is added to the system. The exact details of what must be done, and the order in which the steps are taken, are a function of the particular system. This section is intended only to illustrate what must be done. The frame of reference is a TDM/TDMA system, in which the downstream data consists of data supporting, typically, 24 to 30 lines simultaneously.

Marshaling and Provisioning

Acquisition of a new NID presents some interesting problems. When a new NID is added in the field, it must first be *marshaled*. This process includes being recognized by the headend system and being assigned a frequency and level at which to transmit. After marshaling, the NID is able to communicate with the headend, but it is not necessarily enabled to handle telephone calls. It must be *provisioned* by receiving software and/or setup parameters that configure it for the assigned task set.

Frequency Assignment

The first step in marshaling is to communicate the frequencies on which the NID is to receive and transmit. The NID searches its entire receive band until it finds a downstream frequency on which it recognizes data. One of the pieces of information the NID will be listening for is the upstream frequency on which it is to transmit. The initial transmit frequency may be the one to which the NID will eventually be assigned, or it may be a special marshaling frequency, which we explain later.

TDM/TDMA Summary

Although not all systems will use TDM/TDMA for transmission on the cable, many early entry systems are using it. This section summarizes and illustrates some of the concepts described previously.

Figure 6.11 illustrates several significant aspects of a typical TDM/TDMA signal. The downstream TDM signal is illustrated in Figure 6.11(a). In this example, it consists of 24 time slots, each of which can accommodate one telephone call in the downstream direction. The frame starts with a header, which marks the start of the frame. The frames are 1 ms long, for a total of 1,000 frames per second. Since each telephone call is sampled 8,000 times each second, each time slot must contain a packet of 8 samples, each 8 bits in length. In addition, the time slot must include overhead for operating the NID. Instructions to ring the phone, supply dial tone, adjust upstream level, reconfigure (reprovision) itself, accept a new software download, and other instructions must be sent to the NID. Because of this, a protocol that supports 24 lines will likely

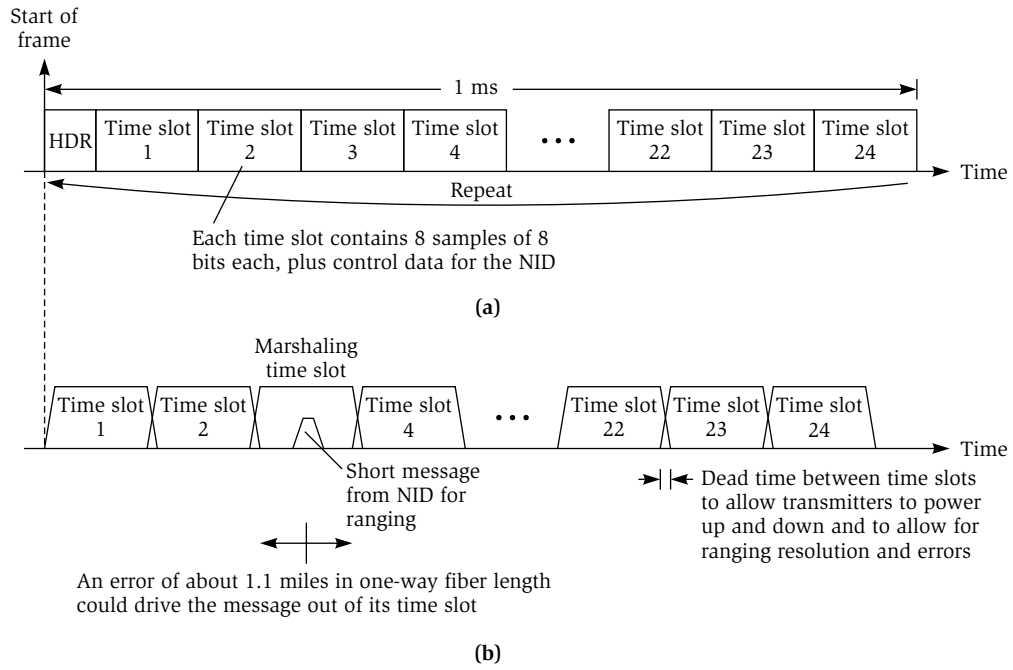

Start of frame
1 ms
HDR
Time slot 1
Time slot 2
Time slot 3
Time slot 4
Time slot 22
Time slot 23
Time slot 24
Time
Repeat
Each time slot contains 8 samples of 8 bits each, plus control data for the NID
(a)
Marshaling time slot
Time slot 1
Time slot 2
Time slot 4
Time slot 22
Time slot 23
Time slot 24
Time
Short message from NID for ranging
Dead time between time slots to allow transmitters to power up and down and to allow for ranging resolution and errors
An error of about 1.1 miles in one-way fiber length could drive the message out of its time slot
(b)

Figure 6.11 TDM and TDMA transmission. (a) Downstream TDM signal stream. (b) Upstream TDMA signal stream with time offset measurement.

operate at a higher data rate than 1.544 Mb/s, the DS1 rate described earlier, which supports 24 lines. The next standard rate available is the E1 rate of 2.048 Mb/s.

Note that in a TDM system, it is not necessary to allow time between individual packets (time slots). The header defines where each bit occurs between it and the following header. It is necessary only to count the number of bits that have been received since the header to know which bit is now being received.

Contrast this to the TDMA datastream used in the upstream direction (see Figure 6.11(b)). Packets from individual return transmitters "parade" back to the headend. The same packet structure as in the downstream data may be used, but it is necessary to allow some small time between the packets. This time allows one transmitter, which has sent its packet, to turn off, and the next transmitter to turn on and stabilize before it begins transmitting its packet. Because of the need for dead time, it is necessary either to transmit at a higher instantaneous bit rate during a packet or to not transmit some bits in the upstream direction.

Shown in the TDMA signal is a marshaling time slot, which may be reserved, at least at times, for marshaling a new NID. Some options for marshaling the NID were discussed earlier. In one possibility, a new NID transmits in the time slot reserved for marshaling. The problem is that the range (signal transport delay) from the NID to the headend is not known initially. If the assumed range is in error by as little as 1 mile or so of fiber, it is possible that the response from the marshaling NID will fall outside the allocated time slot, interfering with another telephone conversation.

Time Slot Assignment

The switch interface, modem logic, and modem bank controller of Figure 6.10 work to implement a number of functions required by the HFC system. Time slots (in a time division multiplexed system) are assigned to each telephone call. In early systems, this was often done on a permanent or semipermanent basis. Later systems include concentration, in which time slots are assigned dynamically as needed.

NID Identification

It may be that when a NID is issued from the warehouse, it is identified to the modem bank, which begins transmitting its address in a downstream datastream. This downstream frequency may be a separate acquisition frequency, or the new addresses may be included in an operational datastream. Alternatively, a "successive approximation" technique may be used. In successive approximation (so named from the operation of some analog-to-digital converters), a hailing signal is transmitted that says, for example, "all unmarshaled NIDs whose address begins with a 1 reply now." If no replies are received, the headend knows that no unmarshaled NIDs have addresses beginning with 1. It can then ask for NIDs with addresses beginning with 0. A variation of this protocol doesn't initiate the marshaling process unless it is determined that an unmarshaled NID exists on the node. This could be determined by reserving a time slot in which any unmarshaled

NID will transmit. The headend listens for any reply during the time slot. When it hears a NID reply, it begins a successive approximation search to determine the address of that NID. In the unlikely event that more than one NID is unmarshaled and replies, the successive approximation search will find all of them eventually.

NID Transmitter Level

The HFC protocol must manage the level of the transmitter in the NID. As shown in Chapter 16, level control is crucial to proper operation of the cable plant. The most critical need for tight level control is at the input to the upstream laser. However, since it is located in the field, the best the system can do is to observe the return signal level in the headend and command the NID transmitter to increase or decrease its level until the level in the headend is correct. In order to know what is "correct," the engineer must know the gain between the upstream transmitter input and the point in the headend at which gain is measured.

The headend system must continuously monitor the upstream level it receives from each transmitter and must signal the individual NID transmitters when they need to adjust their level up or down. Level resolution of less than 1 dB is recommended. The resolution with which the NID can be set should be somewhat finer than the dead zone of the headend level tolerance. This will prevent continuous "hunting" for the correct level. The headend system must accurately measure the level of each return signal. Some early systems have been reported to measure bit error rate (BER) and to command the NID transmitters to increase level when the BER drops. This is not correct because if the BER is compromised from interference that has developed on the frequency, all this algorithm will do is to overdrive the return laser transmitter. If the NID cannot communicate with acceptable BER and with nominal signal level, it must be moved to a new frequency.

Implicit in any acquisition protocol is the need to get the headend to recognize the presence of the new NID. The NID must transmit to make this happen, but it has no a priori knowledge of the level at which it must transmit. It cannot risk transmitting at too high a level because doing so will cause distortion, and probably clipping, of the upstream laser. Other services using the return path will experience interruption. The rational protocol would be one in which the NID transmits at its lowest level and waits for a reply from the headend. Hearing none, it increases its level somewhat and transmits again. The process continues until the NID gets an acknowledgment from the headend. At that time, the headend can command the NID to the correct level.

Timing Offset

After the NID is recognized by the headend and its transmit level is established, it is assigned a time slot on a TDMA upstream frequency. The time slot may be permanently assigned, or it may be assigned only for the duration of one phone

call, then reassigned to another call. Not only must the time slot be assigned, but a time offset value must be assigned, based on the distance from the headend to the NID. The round-trip signal delay from the headend to the NID and back must be measured, and the NID must be given a time offset, which tells it how much to advance its clock with respect to the headend clock so that when it transmits upstream, its signal will be received at the headend at the proper time.

The speed of light (or radio waves) in free space is 186,000 miles per second. In coaxial cable, it is typically 80% to 90% of this, depending on construction of the cable. Fiber-optic cable is denser, so the velocity of propagation is around 60% of that in free space, or around 111,600 miles per second. If 10 miles of fiber are used to reach a node, then the signals will experience a round-trip delay equivalent to 20 miles of fiber. The delay is given by

$$delay = \frac{round\text{-}trip\ distance}{velocity\ of\ propagation} = \frac{20}{111{,}600} = 179\ \mu s.$$

At a data rate of 2.048 Mb/s, 1 bit is transmitted in 0.45 μs, so the delay time is equivalent to about 367 bits. Obviously, the propagation delay must be compensated. This compensation can be done by measuring the time delay of a sample signal, then computing an offset time, which is transmitted to the NID. The NID synchronizes its transmission to a "time = 0" marker transmitted from the headend. It counts the number of clock cycles from this marker to determine its transmit time. The number of cycles is offset based on the measured round-trip delay from the headend.

In order to make the time delay measurement, the NID must transmit a message at a time known to the headend modem logic. The modem logic measures the delay in the time the return signal was received and compares that with the time it expected the reception. The difference in the two is the time advance that must be transmitted to the NID. This measurement may be done on the operating channel, where other NIDs are already operating. The problem is that the initial transmission from the NID must not interfere with communications from another NID transmitting in an adjacent time slot. In order to overcome this problem, the system may initiate a very short transmission in the center of the time slot being used for marshaling. This will limit the fiber distance over which marshaling may be accomplished with no a priori knowledge of that distance to about 1.1 miles. (This is based on a TDMA system having 24 time slots, 1,000 frames per second, and a timing error of one-half of one time slot.)

The distance may be increased by telling the modem logic how much fiber is in the nodal path so that it does have a priori knowledge of the distance. Then only the much shorter delay in the coax portion need be considered. Another way is to use out-of-band marshaling to achieve crude distance measurement, which is then refined if necessary after the NID is moved to its operating frequency.

6.4.6 Implication of Diverse Routing

In order to improve the availability of cable telephony systems, some operators are installing redundant node receivers and transmitters, and connecting them to the headend using fiber-optic cable having diverse routing. Figure 6.12 illustrates use of a redundant node, which has two optical receivers and two optical transmitters. (Chapter 18 includes more information about plant architectures.) Normally, the selection switches are in the positions shown, coupling the primary receiver and transmitter to the headend. If the received signal is lost, this fact is detected, and

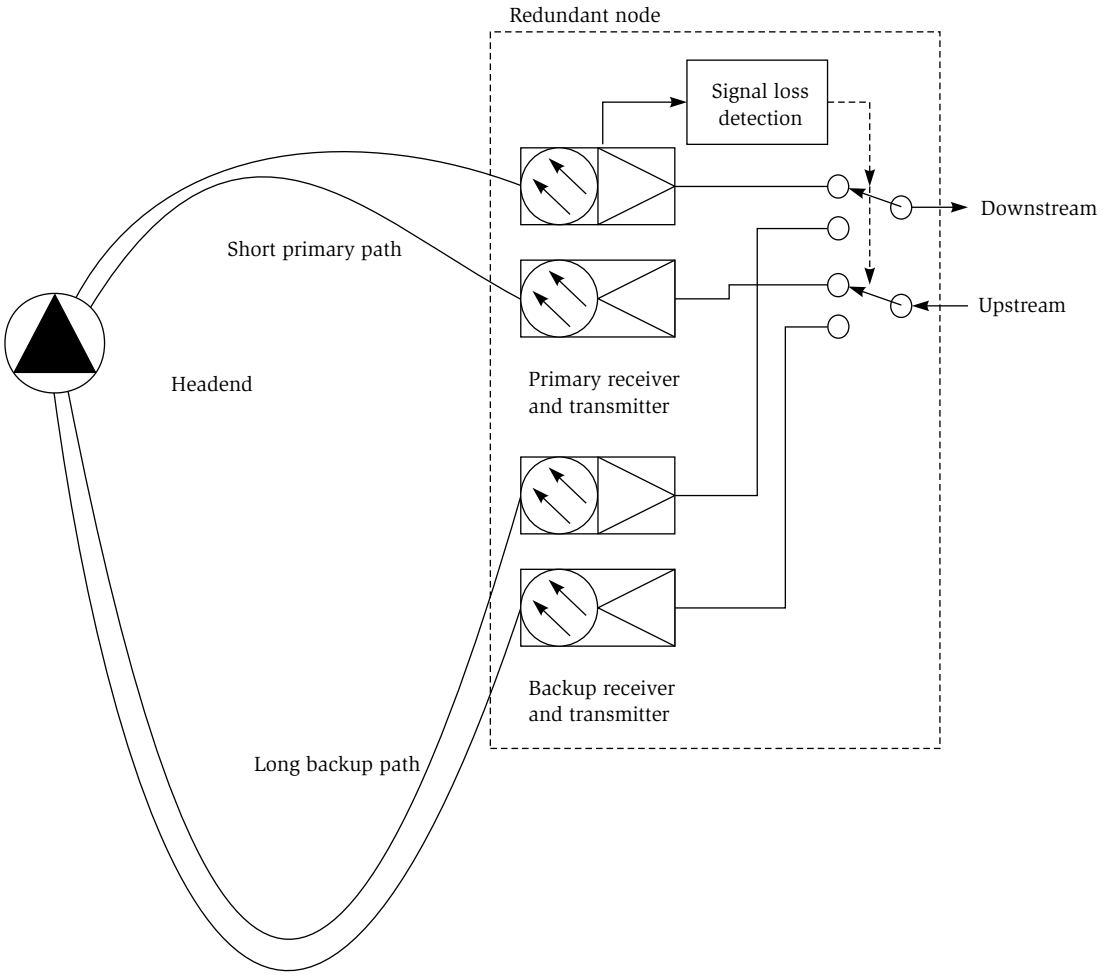


Figure 6.12 Redundant node with diverse fiber routing.

the switches thrown to the opposite position, to couple to the backup equipment. The loss of signal may be based on either loss of light or loss of RF signals. Normally, it is assumed that if downstream light or RF is lost, the reason is a cut in the fiber cable, so the reverse path is also affected. For this reason, most redundant systems cause the upstream switch to follow the downstream switch.

Typically, a telephone system will tolerate loss of the link for much more than 50 ms before the call is dropped, so it is an accepted practice to detect the failure and effect switching in less time than this. The backup path is almost certainly a different length from that of the primary path. This means that the marshaling of all NIDs will be wrong when the backup path is selected. The problem is only a translation in time if all the NIDs are similarly affected by the change in propagation delay. However, it is possible to split a headend telephone modem, with some of its NIDs in one node and some in another. If this has been done, then complete remarshaling will be necessary.

One proposed solution is to build two marshaling tables into the modem logic and to notify it when the backup route is in use. This means that, at some time before the backup route is needed, it will have to be activated and a second marshaling table built. If changes are made in the backup routing, then the process will have to be repeated. If the same modem serves more than one node, then backup tables must be implemented for each node served, and the modem logic will have to be notified of which node has transferred to its backup path.

6.4.7 Alternative Transmission Methods

Most of the preceding discussion was based on TDM/TDMA transmission protocols. These are currently in widespread use, but other methods have been introduced to the marketplace as well. Alternatives include SCPC, CDMA, and PFDM. They and more are covered in Chapter 4.

6.5 Network Engineering: Quality of Service

The art of network engineering deals with providing adequate equipment such that subscribers don't experience undue *call blocking*. A call is said to be *blocked* if it cannot be made because of lack of network resources. Provision of adequate equipment must be weighed against the need to control equipment costs by not providing more equipment than needed. A fundamental concept of network engineering is the busy hour, the hour in the average day at which traffic peaks. This usually occurs in late morning. If the network was engineered to carry the peak traffic on the busiest day of the year (high day busy hour, HDBH, usually Mother's Day), then at all other times the network would carry traffic with no call blocking. However, the cost of the network would be prohibitive. Most of

the equipment would not be used most of the time. Balancing the need for low call blocking and low equipment cost is a domain of network engineers.

The grade of service, or probability that a call will be blocked, assigns a percentage of time (0% to 10%) that calls are allowed to be blocked when the network is congested. A key variable is the amount of time a line is being used. Planners use, as a basic unit of measure, the centum call second (CCS). One CCS is 100 seconds of line use per hour. For example, a 3-CCS line is in use, on the average, for 300 seconds in that hour. A related term is the *erlang*, the maximum physical capacity of a line or trunk. One erlang is equal to 36 CCS, or 3,600 call seconds (or one call lasting for an hour). A 3-CCS line is thus utilized to $3/36 = 0.0833$ erlang.

A common assumption of utilization of a residential line is 3 CCS (equivalent to usage for five minutes an hour). Business lines are assumed to be 6 CCS, and an ISDN line 12 CCS. With the emerging popularity of second residential lines, often used for Internet access, the usage of residential lines may be increasing. Network engineers estimate usage of a line, and combine this with Poisson distribution tables, or similar tools, for predicting when a call is likely to be initiated and how long it will last. This information is used to predict the quality of service or the relative infrequency that a call will be blocked. Adequate equipment is provided to obtain the desired quality of service.

6.5.1 Call Concentration and Blockage

In order to achieve optimum performance in terms of cost and service level, it is important to *concentrate* calls. That is, since not every telephone is used all the time, the time slots on the HFC plant and the time slots in the interface with the switch may be shared between users. Design of concentration is part of the job of the network engineer. The complete job of network engineering is beyond the scope of this chapter, but information should be available from the chosen telephony vendor. Only the basic principles are presented here.

In HFC telephony, concentration may be applied independently to two different parts of the network. The interface between the HFC telephony system and the switch (if it is a TR-303 interface) can be concentrated. This path is typically shared by a number of headend modems, so the universe over which concentration can be achieved is large.

The other point of concentration is in the RF plant. It is not necessary to assign a permanent ("nailed up") time slot to each line, but rather, the time slot each line occupies can be assigned on a call basis. By allowing all lines to have access to the entire universe of time slots, significant *RF concentration* may be achieved. When a line needs a time slot (that is, is involved in a phone call), any time slot not previously in use can be assigned to it. Lines not currently active in making a phone call are not tying up a time slot.

Figure 6.13 illustrates call concentration. The concentration group illustrated is too small to be of practical significance, but it illustrates the concept. Each call is represented as a rectangle, with time on the horizontal axis. The first

number in each call identifies the telephone: telephones 1 through 5 are involved. The number after the dash is the sequential call number from that telephone. Telephone 1 made two calls, telephone 2 made one, and so on.

In Figure 6.13(a), each telephone is assigned its own time slot (or spreading code, or frequency, or copper pair, or whatever resource is to be shared). Concentration is not practiced in this case. Since five time slots are dedicated, the total capacity is 5 times 36, or 180 CCS, the resource dedicated to handling calls from these five phones. The total call time occupied is 68.4 CCS, so the efficiency is $68.4/180 = 38\%$.

Figure 6.13(b) illustrates what happens if only three time slots are dedicated to these five telephones. Note that call 3-1 was assigned to time slot 1, the only

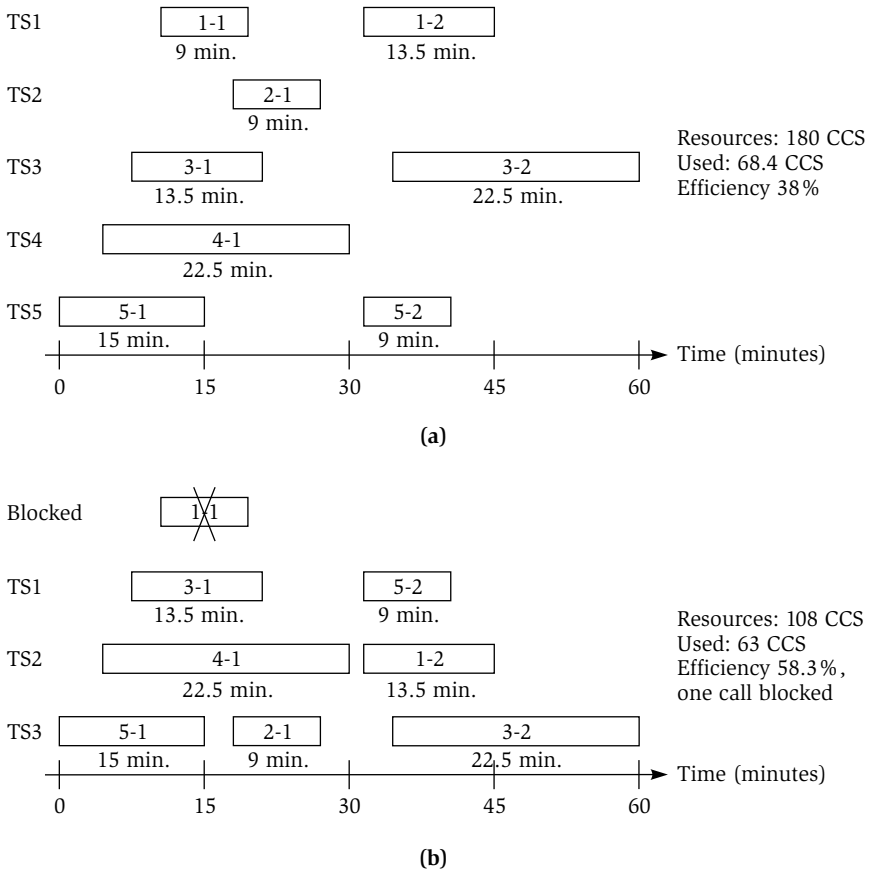


Figure 6.13 Call concentration. (a) No concentration. (b) 5:3 concentration.

slot open when that call was initiated. When call 3-2 was initiated, it was assigned to time slot 3. Call 5-1 was assigned to time slot 3, but when call 5-2, involving the same phone, was placed, it was assigned to time slot 1. In this example, since only three lines are dedicated to serving these same five phones, the total resources tied up were only 108 CCS, a savings of 40% in resources. Because of this reduction in resources, the efficiency with which the resources were utilized increased from 38% to 58.3%.

The penalty, though, is that call 1-1 was blocked. When it was initiated, all three lines were busy, so instead of a ring, the caller got an all-trunks-busy signal (see Table 6.2) even though the phone was free. Of the eight calls considered, one was blocked, so the blocking rate was $1/8 = 12.5\%$.

Real concentration groups should be larger than that illustrated in Figure 6.13 in order to provide the best concentration. The larger the concentration group, the better the concentration because statistics are working for the network engineer: the more lines available, the greater the likelihood that one of them will be available when needed. Figure 6.14 illustrates this principle.[7] This curve plots the percent blockage accepted on the horizontal axis and the percent of maximum capacity utilized on the vertical axis. Percent of maximum capacity is computed by dividing the number of CCS actually carried by the number of CCS of capacity represented by the number of channels available. In order to convert this to the number of subscriber lines, the network engineer must determine the average CCS load offered by each line.

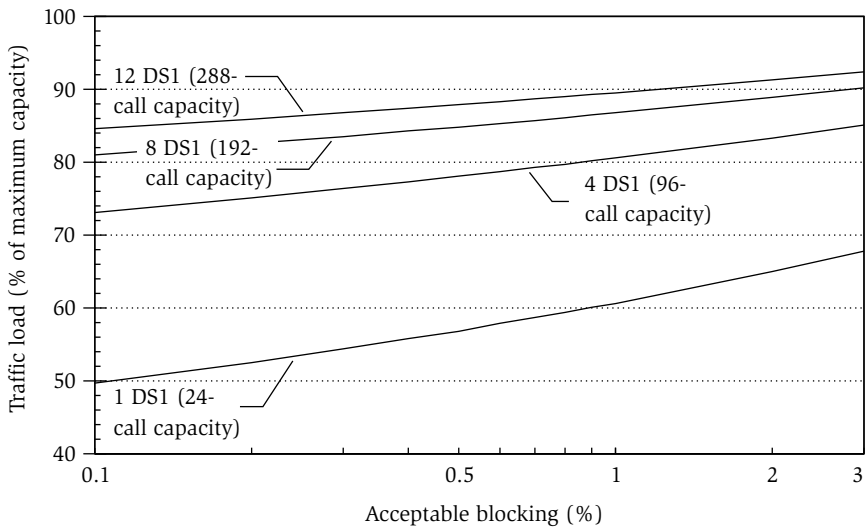


Figure 6.14 Traffic load versus blocking for differing capacities.

If the concentration group is served by only one DS1 (capable of 24 simultaneous conversations), then, depending on the blocking tolerated, the efficiency of utilization of that resource ranges from 50% to 68%. On the other hand, if the concentration group is large enough to support 12 DS1s (288 simultaneous calls), then the efficiency will range from 85% to 92% for the same range of tolerated blocking. The reason is that, as the number of lines increases, the probability of blocking goes down simply because there are more possibilities of finding an unused time slot somewhere in the group.

This efficiency speaks toward designing larger telephone nodes. Large nodes introduce other problems, however. First, large nodes allow more noise concentration, reducing the probability that the reverse path will work properly. Second, the larger the node, the more reverse spectrum will be required to operate the telephone system. A typical practical number for first-generation telephony systems is that a group of 24 to 30 channels will require about 2 MHz of both upstream and downstream spectrum.* In the preceding example, if only a single DS1 is required to service a telephony node, then 2 MHz of reverse spectrum will suffice (not including backup spectrum for use if the primary spectrum becomes corrupted). On the other hand, if 12 DS1s are used, then 24 MHz will be required. With some spectrum held in reserve (to allow for ingress), this could easily consume all the usable reverse spectrum. Network engineering on the HFC network involves trade-offs in the size of the nodes, the tolerated noise, the spectrum allowed for telephony, and the quality of service (percent blocking tolerated).

6.6 IP Telephony

IP telephony began when several proprietary computer-based voice communications systems used the Internet to avoid long-distance charges. Voice signals were digitized, frequently in a sound card, compressed in software, and sent as data packets over the Internet. Due to practical limitations in the computer, *half duplex communications* were employed. (In half duplex communications, only one side may speak at a time; when it is time to allow the other side to transmit, something must be done to enable the reverse transmission.) Later these early IP telephony systems were enhanced with the addition of gateways at which the IP telephone call could be connected to the public switched telephone network (PSTN).

---

* If only a 1.544-Mb/s DS1 were carried, the theoretical RF bandwidth for QPSK modulation would be just over 0.75 MHz. As a practical matter, the data rate for 24 calls must be somewhat greater than 1.544 Mb/s to allow overhead for control signals. Further, as more excess bandwidth is used, the reliability of communications improves. Thus, though a smaller bandwidth is possible, 2 MHz has been used by several manufacturers as a basis for one "set" of 24–30 voice channels. Chapter 4 covers the bandwidth issues in more detail.

IP telephony has now been recognized as the next major change in the way telephone traffic is handled. The name *voice over Internet protocol* (VoIP) has been given to the technology. No longer seen as a way to bypass the long-distance network, VoIP is seen as offering features and efficiencies compelling enough that it will gradually supplant switched circuit telephony, though the transition will take many years. VoIP has been slower to develop than some expected, however, due to a number of practical problems that must be overcome.

6.6.1 A Simplified Description of VoIP

Figure 6.15 offers a simplified comparison of the functionality of a switched circuit telephone system and a VoIP system. The VoIP system shown is what is known as a *client-server* system, the type of system that the cable industry is installing. There is another type of VoIP system, known as a *peering* system, which will be described briefly later.

In the switched circuit system of Figure 6.15(a), the switch handles all call functions and also the signals to and from the end telephones. As described earlier in this chapter, the telephones may be connected either directly to the switch or through a digital loop concentrator (DLC). In either event, a dedicated pair of wires is assigned to each phone number from the switch or DLC to the phone.

As an example of a switched circuit call, consider a call placed by subscriber A to subscriber B. A picks up the phone, which notifies the switch to place a dial tone on the line. A then dials B's number. The switch has to know which of thousands of wire pairs, or *circuits*, is B's connection. It then places a ringing voltage on B's line and a ring-back tone on A's line. When B picks up the phone, the switch removes the ringing voltage on B's line and removes the ring-back from A's line. It then completes a voice circuit in both directions from A to B and B to A and maintains this connection for the duration of the call.

Now consider the client–server VoIP system shown in Figure 6.15b. Several devices are involved in completing the call, as will be explained in more detail later. The softswitch (server) sets up the call, but then it is not involved again until the call is completed. The router shown is the same router that's handling other data packets to and from customers. There is no separate wire pair connecting each subscriber's telephone to a central point. Rather, the signal is carried on coaxial and fiber cables, along with data and television signals.

The customer gateway may consist of a cable modem plus specialized circuits (the client) to connect to the telephone. Data from that subscriber's client is multiplexed with other data from the subscriber plus data from other subscribers and sent to the headend. The telephone signal is multiplexed with other data, frequently but not always using TDM/TDMA (see Chapter 4). This is the first point of simplification for VoIP, as compared with switched circuit telephony. No separate circuits are used to connect the subscriber to the central office (in fact, in classical terms there is no central office). Rather, the telephony

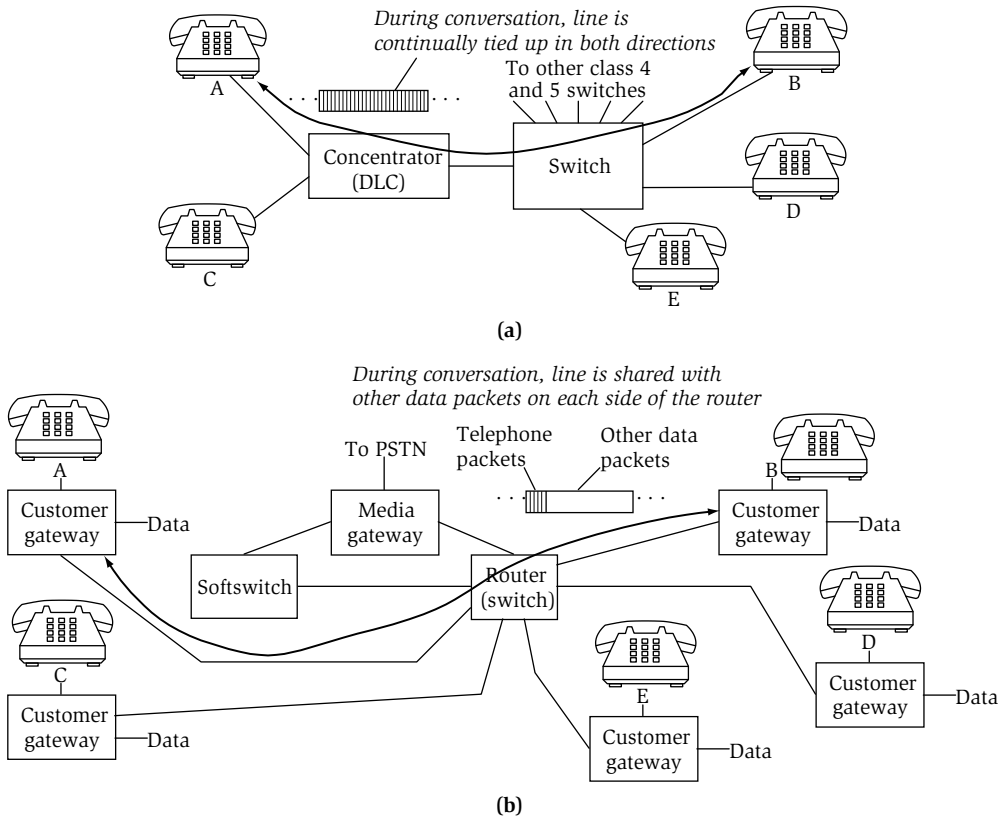


Figure 6.15 Simplified comparison of switched circuit (a) and VoIP (b) telephone systems.

data traverses the same circuit as does all other data. The inherent multiplexing capability of cable modems drastically reduces the number of router ports needed.

When A wishes to call B, A picks up the phone. This sends a packetized signal via the router to the *softswitch* (the server), which takes on some of the functionality of a normal telephone switch. If the subscriber is authorized, the softswitch sends a packet telling the customer gateway (the client) to apply a dial tone to A's telephone. A dials B's phone number, which is sent to the softswitch. The softswitch obtains the IP address for B and sends that address to A's customer gateway. The softswitch then sends a packet to B's gateway, telling it that A is calling. B's gateway applies a ring voltage to B's phone and provides caller ID if that service is included. At the same time, the softswitch

instructs A's gateway to play ring-back to A's phone. When B answers, B's gateway removes the ringing voltage and sends a packet informing the softswitch that B has answered. A and B then communicate directly by sending voice packets back and forth over the IP network represented by the router. Note that the softswitch is not involved in the voice call itself and that A and B never communicate directly with each other during call setup. The softswitch retains information that the call is in progress, so someone else trying to call either A or B will get a busy signal.

The phone call progresses as each end sends packets of voice signals to the other end. The voice signals are usually embedded in IP packets using a suitable layer 4 protocol, such as *User Datagram Protocol* (UDP) (see Chapter 5). The phone call packets are multiplexed along with all other packets to and from a number of customers, using the same multiplex structure as for all other data. The router is already needed to serve data purposes, and the telephone call is simply a few additional packets that must be transported.

The softswitch is nothing more than a big database, which must have up-to-date information on the phone number and identification of all subscribers. Whereas a class 5 switch is located in a neighborhood and typically serves a few thousand subscribers, the softswitch, depending on its design capacity, can serve many more subscribers and may be located at any convenient point in the network. Once the call is established, the softswitch's job is done, and it can go on to set up another call. Call packets do not flow through it; rather, they flow through the router, as do all other packets. There is no traditional switching fabric; instead, the switching fabric is replaced by routing within the network of routers that is provided for other data services anyway.

Of course, real systems are somewhat more complex than this simplified picture. Most calls don't originate and terminate on the same switch (switched circuit), nor are they carried through only one router (VoIP). The VoIP system must interface to the PSTN to transfer calls not destined to another telephone on the same system, and both must connect to long-distance switched circuit trunks today in order to complete long-distance calls. While it is possible to complete long-distance calls via the public Internet, this is not a common practice, due to quality of service (QoS) issues, which are discussed in Chapter 5.

6.6.2 A Practical VoIP System

Figure 6.16 illustrates a practical VoIP system based on the *Multimedia Gateway Control Protocol* (MGCP).[8] Where several names are in common use for the same block, we have shown all of them. The term frequently used in the PacketCable specifications is italicized (there is some inconsistency in the terminology used in the PacketCable specifications). MGCP is a popular protocol for VoIP systems. CableLabs has defined a subset (plus extensions and modifications) to be the standard PacketCable protocol for signaling between network platforms, called the *Trunking Gateway Control Protocol* (TGCP). Another subset of MGCP, that

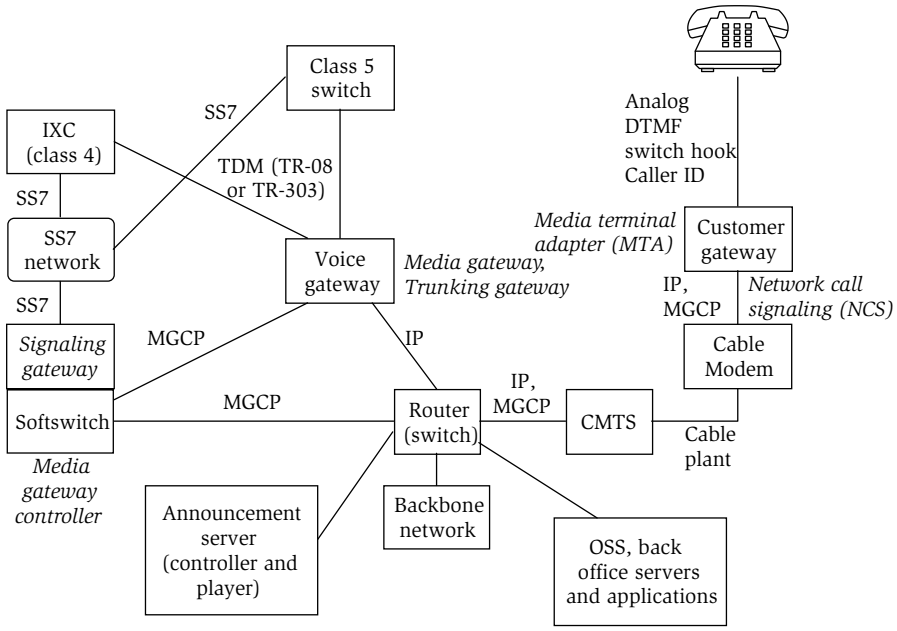


Figure 6.16 Practical VoIP system based on MGCP.

part used for the signaling between the subscriber's phone and the softswitch, was adapted and named *network call signaling* (NCS). NCS is used with TGCP.

Referring to Figure 6.16, start with the telephone in the subscriber's house. It communicates with a *customer gateway*, the client, which contains the conversion equipment to translate signals to the NCS format used on the cable plant. Customer gateways and other points at which the voice packets can end are known as *endpoints*. The gateway may be built into an internal modem, or it may reside on the outside of the house, with a DOCSIS 1.1 or higher cable modem (if used). The telephone uses the common analog interfaces described earlier in this chapter, including analog voice, DTMF tones, switch hook status, and caller ID functions. Among other things, the customer gateway will contain interface circuits, often called a *subscriber line interface circuit* (SLIC) and a *subscriber line audio-processing circuit* (SLAC). The SLIC provides the battery feed, overvoltage protection, ringing, supervision, coding, hybrid transformer, and testing (BORSCHT) functions for the telephone. The SLAC processes voice-band analog signals into pulse code modulated (PCM) outputs and converts the PCM inputs into analog outputs.

The plant-side interface from the gateway in the example is an Ethernet interface bearing IP packets, with MGCP or TGCP as the call management protocol.

(Ethernet is used to connect to the modem. If the customer gateway were integrated with the modem or other network interface, Ethernet would not be needed.) The customer gateway may employ voice compression to reduce the bandwidth, but compression is not advisable from a quality perspective. (See the later section on compression in voice circuits). The cable modem connects to a CMTS at the head-end, and in turn the CMTS interfaces with the router facility. One of the devices connected to the router is a server called a *softswitch*, or, in PacketCable parlance, a *media gateway controller*. This is the device described earlier, which contains the intelligence of the VoIP system. It has a database associating phone numbers with *endpoint names* for subscribers on the network. Endpoint names are constructed in the same manner as are email addresses. Of course, there must ultimately be a translation to an IP address to permit delivery of packets. It is possible to associate phone numbers with IP addresses, but using endpoint names enhances the reliability of the network by allowing multiple softswitches to be available and to concentrate the IP address database in one place (the DHCP server — see Chapter 5).

For subscribers residing on the VoIP network, the softswitch instructs them to send voice packets to each other; the softswitch then has nothing further to do with the call except to be informed when it ends. The packets carrying voice communications between callers are collectively known as the *bearer channel*. Contrasted to the bearer channel is the *signaling channel*, or *signaling plane*, which is the collection of packets used to set up and tear down a call.

For subscribers not on this network, the softswitch knows the endpoint name or IP address of at least one *voice gateway*, the device that interfaces between the VoIP system and the PSTN's class 5 switch and with an inter-exchange carrier (IXC — a long-distance company). As described earlier, when a phone on the system goes off-hook, the softswitch is notified and instructs the customer gateway to send a dial tone to the telephone. The softswitch receives the dialed number and associates it with either an endpoint name on the VoIP network or, if the called party is not on the network, refers the call to the voice gateway, which converts it to a standard switch interface, such as TR-303 (described earlier in this chapter). At the same time, the softswitch provides calling instructions to either the PSTN class 5 switch or the IXC class 4 switch, giving it the called party's number. The signaling gateway may be a separate unit, but often it is incorporated into the softswitch and provides translation to the SS7 signaling protocol used by the switches. Frequently, the functions of the softswitch, signaling gateway, and voice gateway will be incorporated into one device, also called a softswitch. Sometimes the functions are built into a class 4 or class 5 switch, providing an integrated system for smaller operators.

A gateway may be used to interface to a class 5 switch, with the class 5 switch performing the functions of the softswitch. The gateway in this case must be able to translate SS7 signaling into the appropriate VoIP commands. A soft-switch would not be used in that configuration. More capability is demanded of the class 5 switch than if it were used just to handle calls originating or terminating off the VoIP network.

6.6.3 Key Concepts of MGCP and TGCP

The material presented here applies generally to both MGCP, the IETF standard, and to NCS/TGCP, the PacketCable standard. Although we shall normally refer only to MGCP, we really mean both. Differences are noted where they exist.

MGCP assumes a connection model where the basic *constructs* (concepts) are *endpoints* and *connections*. Examples of endpoints include customer gateways, voice gateways, and softswitches. Note that several endpoints can exist in one product, such as a softswitch combined with a voice gateway, which likely has software supporting other services, such as interactive voice response. *Connections* connect these endpoints as required for a call. Connections are grouped to form a call. One or more connections can belong to one call. Connections and calls are set up at the initiative of one or several *call agents*, the softswitch software that sets up and manages the call. If the same softswitch handles both voice gateways involved in the call, there is usually only one call agent at work. If the two voice gateways are handled by different softswitches, then the two call agents communicate between themselves to set up the call.

A key feature of MGCP, as opposed to certain other VoIP protocols, is that the customer gateways have relatively little intelligence. The intelligence resides at the softswitch as much as possible. This has several benefits. The first is that the cost of the customer gateway can be lower, since it does less. Also, features may be added to the system by merely adding software to the softswitch and not doing anything to the customer gateways.

For example, the softswitch instructs the customer gateway to notify it when a phone goes off-hook. When so notified, the softswitch instructs the customer gateway to collect dialed digits. In MGCP, the customer gateway may be instructed to send each digit to the softswitch as it is sent by the telephone. But this generates a lot of short messages on the network, and many short messages are inefficient in their use of the network. A single long message is less wasteful. The softswitch may instead instruct the customer gateway to collect all dialed digits before sending them to the softswitch. This creates a complication for the customer gateway, though, in that different numbers of digits may have to be dialed, depending on what the subscriber is dialing. A local number in North America is dialed with either seven or 10 digits, depending on the location, whereas a domestic long distance number is dialed with 11 digits, and a call to outside of North America is dialed with 13 or more digits. This can be resolved by having the softswitch supply the customer gateway with a *digit map* describing how many digits are expected, depending on the first digits dialed. Digit maps are not a part of the present NCS/TGCP specification: Each digit is forwarded to the softswitch as it is dialed, in short packets. This simplifies the operation of the network, at a slight expense of inefficiency.

The softswitch will also instruct a customer gateway to ring the phone, to provide a ring-back tone (telling the caller that the called phone is ringing), and the address(s) to which the customer gateway is to send packets. MGCP allows

for a number of different networks over which packets could be sent, but NCS/TGCP assumes an IP network.

Codecs

Codecs (coder-decoders) are the devices that translate between analog voice and the digital signals transmitted on the VoIP network. A number of codecs are standardized by the ITU. The most common codec specifications, and one that all PacketCable and other VoIP implementations must support, is G.711. This codec specifies sampling the voice signal 8,000 times a second and encoding it to 8 bits, using either the mu-law or A-law (see Section 6.3.2). This is noncompressed voice coding, as used in the PSTN. It yields the highest voice quality but demands the highest data rate. The resulting data rate for the payload (bearer channel) is 64 kb/s. Note that this is only the payload data rate, not the transmitted data rate, which is covered later. Absent a compelling reason to use compressed voice transmission, it is highly preferable to use uncompressed voice; when fax or modem traffic is being carried, use of uncompressed G.711 is a must if quality transmission is to be provided. Most systems include the ability to detect when a fax or modem call is in progress and, if compression is being used, to turn it off, reverting to G.711 encoding.

Compression in Voice Circuits

Several compression systems are defined, to be used when it is necessary to minimize the bandwidth taken up by a phone call. G.728 is optional in PacketCable. It compresses voice calls to a payload of 16 kb/s. G.729E compresses voice to a payload rate of 11.8 kb/s and is also supported by PacketCable. G.729A is not supported by PacketCable but compresses to 8 kb/s. G723.1 compresses to 5.3 and 6.3 kb/s, but these are not supported by PacketCable either. Higher-bandwidth systems, such as most fiber-to-the-home systems, tend not to support any compression, because the bandwidth consumed by telephony is small compared to the available bandwidth, and all compression systems degrade quality.

The compression standards just cited use a technology called *linear predictive coding* (LPC) as their compression mechanism. This is a good compression technology to use when handling voice calls, but it tends to break down when handling other types of traffic, such as fax and modem. Here is an extremely simplified explanation: LPC models the human voice track as a "carrier" oscillator that is modulated with a low-frequency signal defining the intelligence to be carried. The incoming voice signal is divided into short segments. For each segment, we attempt to find a frequency that represents the largest portion of the energy present during that frame (the *estimated pitch period* is the reciprocal quantity discussed in G.723). We then construct a time-dependent filter that modifies the resultant waveform so that it fits as closely to the original waveform as possible. The output of this filter is weighted by two filters, a long-term predictor synthesizer and a short-term predictor synthesizer, which improve the correlation between the error

computation and human perception. Finally, the synthesized speech is compared with the original speech, and the mean squared error of the difference is computed. This error is used to improve the quality of the prediction.

The basis for this coding technology is the modeling of human speech as being generated by oscillators called the *vocal cords*, with the sound being modified by a series of filters in the vocal track. Thus, we look at a sample of speech, find a model consisting of an oscillator and filters that match it, and then compare the results of the model with the real thing. The mean error is used to refine the model. We then transmit the parameters of the model to the receiver.

Because the compression seeks a dominant frequency in a segment of speech, it is confused when presented with signals not having a dominant tone, as occur in fax and modem transmission. Inevitably, there are errors in the compressed signal, and as you go to compression that results in lower data rates, the quality of the reconstructed voice drops. Further, more compression means that more processing power is required in the system, increasing costs. Also, as you apply more compression, the delay introduced in the signal worsens, creating problems dealt with later. For all these reasons, it is preferable not to use compression if it can be avoided.

Voice Activity Detection

One method often used with voice compression systems to further reduce the bandwidth of the signal is to detect pauses in speech. Fewer or no packets need be sent when there is no speech. Since it is most common for only one of the two conversing parties to talk at a time, it is possible to halve the bandwidth being used by detecting lack of voice activity and to stop transmitting packets during that time. This is done using a function called a *voice activity detector* (VAD). However, if nothing is sent, you would not hear anything from the other end and would feel that maybe the other party was not there. To overcome this, noise is often injected during periods of no speech. This noise, called *comfort noise*, may be synthesized, or it may be derived from actual background noise, using lower-bandwidth transmission than that used by the actual voice transmission.

6.6.4 Delay

There are several sources of delay in a VoIP network. By delay, we mean the difference in time between when a word is spoken and when it is heard by the listener (or when the echo is heard by the talker, as described later). We already mentioned the delay due to compression. Delay is incurred both at compression and decompression. And there may be additional delay due to the processing circuitry and software. Then the signal must be put in packets. There is a trade-off in the size of the packets. If you put fewer samples in a packet, there is less delay. But since each packet must have a header of 62 bytes (in the example

system — see later), shorter packets are less efficient, in that they "burn" more bandwidth due to more headers. Longer packets have the same-length header; but since more information is transmitted per packet, there is less inefficiency. Longer packets introduce more delay because you cannot transmit the packet until all bits have been accumulated.

Think of the analogy of carrying water in a bucket that you fill from a faucet. The faucet supplies water at a fixed rate (corresponding to the rate at which packets are accumulated at the sending end of the call). If you have a bigger bucket (corresponding to larger packets), you can carry more water at a time (at least until the bucket gets too heavy), so you work more efficiently. But it takes more time (delay) to fill the bucket at the faucet.

Another significant source of delay in IP networks is the need to remove jitter from the packets arriving at the receiver. In order not to introduce uncomfortable interruptions in speech, it is necessary to make sure that each packet has arrived and is ready to be played by the time the previous packet is finished playing. However, IP networks by their nature cannot guarantee a constant arrival time of packets. Sometimes they will arrive after a short delay, and sometimes they will arrive after more delay, depending on network congestion. If you have tried to listen to an audio clip on the Internet, you have likely experienced an interruption in the audio when the next packet is delayed en route to your computer. To avoid any interruptions to the voice, the received packets are *buffered*, or held in queue, at the receiver for long enough to ensure that packets will always be available when needed. The amount of buffering needed depends on the jitter in the network. This *dejitter buffer* is another significant source of delay in a VoIP network.

Delay causes a couple of problems in telephony. If the delay gets too long (several hundred milliseconds), you begin to notice a delay in response from the other end. You ask a question, and the other party doesn't answer as fast as you expect. If you have made overseas calls using satellite transmission, you can appreciate the problem. Some cell phone systems employ enough compression that they tend to approach the point where you would notice the delay. (Try carrying on a cell phone conversation with a colleague in the same room so that you hear each other over the cell phones and directly. You will see how much delay there is.)

One guideline to acceptable delay is found in ANSI T1.508–1998, *American National Standard for Telecommunications — Loss Plan for Evolving Digital Networks*.[9] Section 6.2.1 of that document recommends that the round-trip delay between two end users be kept below 300 ms. To permit this in international circuits, each national segment is allocated 100 ms of round-trip delay. Satellite delays are of necessity longer.

6.6.5 Echo Cancellation

Even before the delay gets long enough to notice a response delay, echo can be a problem. For several reasons, when a signal crosses a boundary where it is

converted between two-wire and four-wire transmission, some signal can be reflected to the talker. This produces an uncomfortable echo. The amount of delay that can be tolerated is subjective, so you might find different authorities using different numbers as the threshold over which *echo cancellation* (ECAN) must be used. ANSI T1.508–1998 suggests that when round-trip delay exceeds 5 ms, an echo canceler should be used. PacketCable suggests that the threshold is 20 ms. It is practically impossible for a VoIP system to introduce less delay, so echo cancellation should be used. The current PacketCable specification requires that all endpoints have ECANs.

Echo cancelers are sometimes as simple as devices that attenuate the far-end received signal, but these are not particularly satisfactory. Most modern echo cancelers measure the round-trip delay and construct adaptive filters that mimic that delay, subtracting the originated signal from the returned signal. Very likely at some point you have started a long-distance or cell phone call and noticed a bad echo, which goes away after a second or two of conversation. This is your echo canceler measuring the echo and adapting itself to cancel it. Each link of a telephone system that has the potential to introduce echo is responsible for canceling the echo it introduces.

There are some conditions, though, under which echo cancellation must be turned off, so all systems have the ability to detect those conditions and to instruct the endpoint to turn off echo cancellation. For example, when a telephone link is handling fax or modem traffic, the type of traffic is detected and the echo cancellation is turned off. Note that even with data and VoIP available, there will still be times when people need to use the VoIP circuit to transmit these older data communications protocols. There is an alternative way to transmit faxes, in which the fax signal is demodulated in the network near the point of origination, the data is transmitted, and the fax signal is reconstructed near the user. This alternative is called *fax relay*, described in ITU T.38. These facilities, not in widespread use, are listed as optional in PacketCable.

6.6.6 Real-Time Protocol

Real-Time Protocol (RTP) is really two related protocols found in many VoIP systems. RTP is used to transmit bearer packets, that is, the packets containing the voice information. *Real-Time Control Protocol* (RTCP) is the related protocol used to carry data relating to setting up and tearing down the RTP session. These two protocols are maintained by the Internet Engineering Task Force (IETF).

RTP — Real-Time Transport Protocol (RFC-1889, -1890). This protocol is used for a number of audio and video services. For VoIP, it indicates the type of compression being used (if any) and includes time stamps. The time stamp is put on at the sender. It is read by the receiver and controls the time that a packet is played. If two packets contain time stamps that are 20 ms apart, the second packet will begin to be played 20 ms after the first packet starts to play.

This is necessary because the packets will not necessarily arrive at constant time intervals. Wallclock time (absolute time) is represented using the timestamp format of the Network Time Protocol (NTP), which measures time in seconds relative to 0 h (zero hours) UTC on 1 January 1900. The RTP header also includes a sequence number, which allows the receiver to know if packets are received out of order and to reorder them if necessary.

RTCP — Real-Time Control Protocol (RFC-1890). RTCP is used to control data packets (the bearer channel). RTP packets are used for the bearer channel, and RTCP is used to communicate housekeeping data between the endpoints. A sender periodically transmits a sender report packet using RTCP. It lets the receivers know what they should have received. Similarly, receivers send RTCP receiver reports, which indicate what the receiver has received. A third type of RTCP packet is a source description, which allows senders to communicate more information about themselves. The information transmitted includes a unique name for the source, a real user name of the source, an email address, a telephone number, a geographic location, the name of the application generating the stream, notes about the source, and private extensions. An RTCP Bye message indicates that a source is leaving a conference. Finally, RTCP can contain application-specific messages. These are intended mainly as an easy way for applications to experiment with new kinds of messages.[10]

Thus, we find two fundamental types of packets in use in the example VoIP system. Bearer channels, containing packets of voice signals, are carried in RTP. Signaling packets are carried in RTCP. The two types of packets do not identify themselves as to which type they are. Rather, this is done through the use of different port numbers at layer 4, where UDP is employed. RTP and RTCP are often, but not always, carried in UDP packets.

6.6.7 User Datagram Protocol

Chapter 5 includes a brief description of UDP. It is an alternate to TCP, used when the packets being sent are not to be acknowledged by the receiver. Acknowledgment is not feasible when a packet is intended for multiple receivers or in cases, such as voice transmission, when retransmission of a missed packet is not workable — by the time the lost packet is identified and a retransmission is received, it is too late to use the packet. There are other methods of covering lost packets, such as repeating the previous packet. UDP packets are usually put in IP packets for transmission on the network.

6.6.8 Makeup of VoIP Packets

Figure 6.17 illustrates the makeup of the IP packets containing VoIP bearer packets according to the protocol suite being discussed, which is used in Packet-

Cable and other VoIP applications. See Chapter 5 for information concerning the makeup of the UDP and IP packets. RTP uses application layer framing, meaning that the application can interpret what is in the packet, so there could be interpretations of the packet other than what we show here.

RTP uses a compact header to keep from adding a lot of overhead. The *version* number defines the version of RTP; currently there is only one version. The *padding* bit indicates whether extra bytes have been added to pad the packet or whether all are valid data. The *extension* bit will indicate if there is any additional header information other than what is shown here; so far, no extensions have been defined. The *contributor count* tells how many contributing source identifiers the message contains. Normally in point-to-point VoIP calls, there will be only one contributing source. However, the contributor count could be used for conference calling, as described later. Next is the *marker* bit, available to the application for its use. Finally, the *payload type* describes the data to follow. A number of different formats of both audio and video may be carried in RTP packets. We are dealing with only a small subset in the VoIP application.

The *synchronization source identifier*, abbreviated SSRC, identifies the original sender of the message. The original sender identifies the *sequence number* and the *time stamp* for the data. Contributing sources are defined when

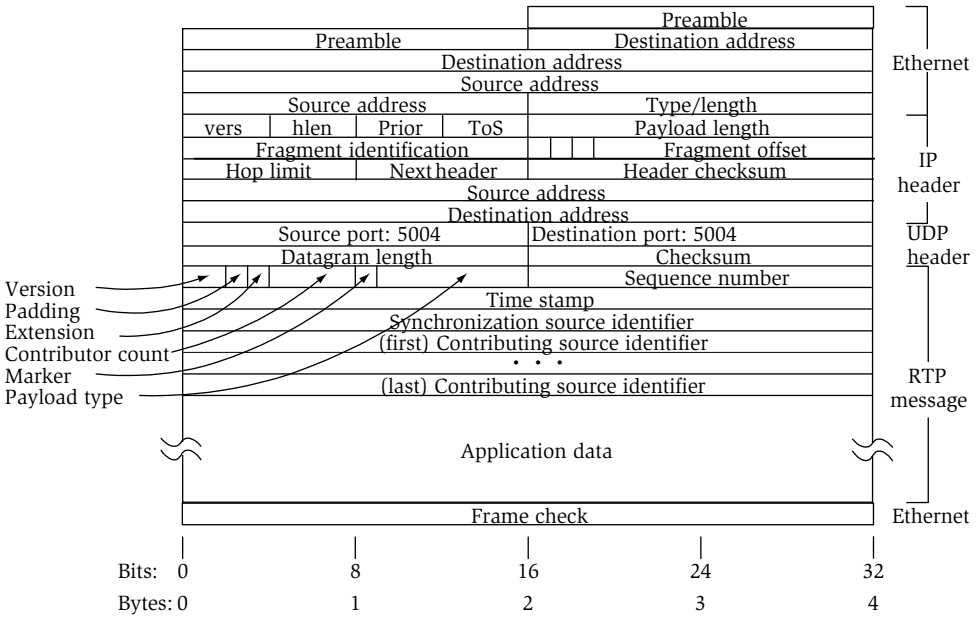


Figure 6.17 RTP in UDP in IP in Ethernet.

more than one endpoint is originating data. There is only one contributing source in a point-to-point voice call, but if you are holding a conference call, you will have more than one contributing source.

6.6.9 Conference Calling

In order to accommodate multiple sources, RTP uses the concept of *mixers*. A mixer serves the same function that an audio mixer performs in audio applications, but it does it in a different way. Applied to the VoIP application (which uses a subset of what RTP is capable of doing), a mixer receives inputs from all the participants on a conference call, adds them, and sends out a new RTP packet to all participants so that they can hear each other at the same time. Note that in applications other than VoIP, a mixer might also combine images or do other operations where it is desired to combine elements from different sources.

Figure 6.18 illustrates use of a mixer to facilitate a conference call between four subscribers. Rather than route packets directly from one subscriber to another, in a conference call all packets are routed to the mixer, which combines them and sends the combined packets to all participants. The mixer may be a

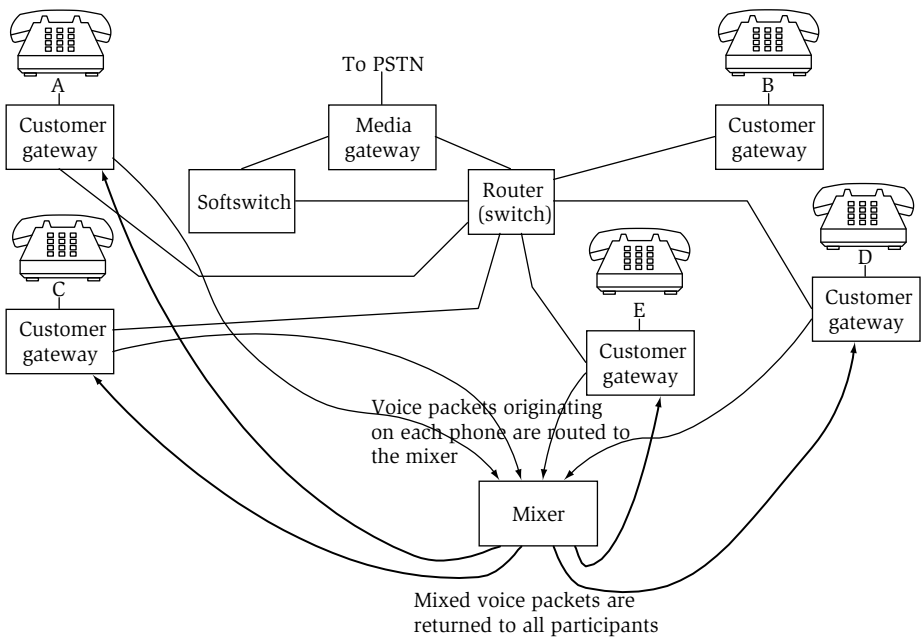


Figure 6.18 Conference calling using a mixer.

separate function, or it may be implemented in the softswitch. It is necessary in a VoIP system only if conference calling is to be supported. Though not supported in early versions of the PacketCable standard, some VoIP systems offer three-way calling without use of a mixer. In these systems, each endpoint is capable of accepting two independent audio streams and mixing them, in essence performing the function of a two-input mixer. In such systems, the mixer would be needed only for conferences of more than three endpoints.

Of course, not all of the call sources will necessarily be on the VoIP network as shown. Some could be on the PSTN and be routed through the media gateway.

While not necessarily useful in typical VoIP networks, we'll mention one other function defined in the RTP protocol, the *translator*. A translator is used to connect a subscriber who has different network capabilities than do other subscribers. For example, suppose one subscriber to a VoIP network was connected using a 64-kb/s ISDN connection, which is not fast enough to support uncompressed voice packets. The subscriber to whom the first subscriber is talking is connected to a cable modem and is not using compression. The translator would receive packets from the cable modem end, compress them so that they fit in the 64-kb/s channel, and send them to the subscriber on the ISDN connection. For packets in the other direction, the translator would decompress the packets and send them on to the cable modem subscriber.

6.6.10 Bandwidth Required for a VoIP Signal

The bandwidth required for a switched circuit signal is normally 64 kb/s, with very little overhead. However, when voice is transmitted in packets, we have to add overhead in order to manage the packets. The required bandwidth thus is the sum of that required to send the voice data and the packet overhead. There is a trade-off between the number of voice samples sent in the packet and the total amount of overhead. The amount of overhead is fixed per packet and is dictated by the protocols being used. A short packet will demand this overhead for fewer voice packets, so the resultant data rate will increase. A longer packet will demand this same overhead for more voice packets, so the total data rate will decrease. The penalty is that more delay is introduced.

Refer to Figure 6.17, which shows the makeup of a typical VoIP bearer (voice) packet carried on Ethernet. The application data is the actual voice data, which in an uncompressed (G.711) system occupies 64 kb/s of bandwidth. It is placed in an RTP packet, which adds a 12-byte header. We are assuming a simple call between two points, so there are no contributing source identifiers. The RTP packet is put in a UDP packet, which adds yet another 8 bytes of header. This UDP packet is placed in an IP packet, which adds 20 bytes (IPv4), and finally the IP packet is put in an Ethernet packet, which adds 18 bytes of header and four bytes of frame check, for a total addition of 22 bytes. Thus, for a single block of application data, we add a total of $12 + 8 + 20 + 22 = 62$ extra bytes of

Table 6.3 VoIP Block Size and Data Rate

Block size, ms	*Bytes of data*	*Total bytes with overhead*	*Blocks per second*	*Data rate*
(a component of delay)	(block size × 8 bytes/ms)	(bytes of data + 62 overhead bytes)	(1,000 ms ÷ block size)	(total bytes × blocks per second × 8)
4	32	94	250.0	188.0 kb/s
8	64	126	125.0	126.0 kb/s
10*	80	142	100.0	113.6 kb/s
20*	160	222	50.0	88.8 kb/s
30*	240	302	33.3	80.5 kb/s

* PacketCable standard.

overhead. These 62 bytes of overhead are added to every block of application data, so the amount of overhead depends on the size of the data block.

As shown in Table 6.3, there is a trade-off between the delay a voice packet experiences and the amount of overhead added. Short delays are good, but they add more overhead by virtue of the fact that more packets must be transmitted to convey the information. In order to mitigate the effect of the added header information, DOCSIS 1.1 and later versions offer a way of replacing the packet header with an abbreviated header, which reduces the overhead. Called the *payload header suppression* (PHS), it is used to suppress packet headers with a short identifier. Rather than transmit the entire packet, a short identifier replaces the header for transmission between the CMTS and the cable modem, which replaces the original header before sending the packet to a non-DOCSIS system.

Because of the time-critical nature of VoIP traffic, it is necessary to provide for good quality of service (QoS) parameters for transmission of voice packets. This may be done in several different ways, as described in Chapter 5. VoIP packets can be transmitted on the public Internet, and this has been done. The problem is that the Internet is not guaranteed to support the necessary QoS, and there is no guarantee that packets will be delivered to the receiver in a timely fashion and in the correct order. In order to handle voice reliably and with quality competitive with the PSTN, it is necessary to make sure that the entire network handling the call has the requisite QoS parameters.

6.6.11 Other VoIP Control Protocols

We have been describing MGCP and NCS/TGCP, common control protocols for VoIP transport in cable TV plants. But there are other protocols you may encounter from time to time.

H.323 is the original VoIP protocol. It and *Session Initiation Protocol* (SIP) differ fundamentally from MGCP in that they place the network intelligence at the endpoints. No softswitches are required, though they can be used. Rather, each endpoint is expected to know the network well enough that it can establish and manage any possible call on its own. Obviously this demands a lot more intelligence in the endpoints, increasing their cost.

MEGACO is a superset of MGCP; it adds new services beyond voice transport management. The fundamental concepts of MEGACO are terminations and contexts. Terminations are streams of data (voice, video, or whatever). They are grouped into contexts, which define two or more terminations that are in communication with each other.

6.7 The 911 System

911 is a standard plan used for emergency calling. A customer anywhere can theoretically dial 911 from any phone and be connected to a nearby emergency dispatch center. The emergency dispatch center can counsel the caller in what needs to be done and can dispatch police, ambulance, or fire services as necessary. In 1967, the President's Commission on Law Enforcement and Administration of Justice recommended that a "single number should be established" nationwide for reporting emergency situations. In November 1967, the FCC met with the American Telephone and Telegraph Company (AT&T) to find a means of establishing a universal emergency number that could be implemented quickly. AT&T announced that it would establish the digits 911 as the emergency code throughout the United States. On February 16, 1968, Senator Rankin Fite completed the first 911 call made in the United States, in Haleyville, Alabama. On February 22, 1968, Nome, Alaska, implemented 911 service.[11]

Today, many 911 systems are "enhanced 911" systems, where the caller's telephone number and address (for a fixed phone) are displayed to the operator along with the name of the telephone number holder and any other information that has been entered into the system. Usually, 911 is run by local county/city government or public safety departments. Some states have statewide 911 systems or mandate 911 at the state level. There is no national governing agency for 911 in the United States.

Cable operators offering primary-line service should be part of the 911 system. But note that if someone is using an IP phone over your IP system, your switching facility (be it VoIP or switched circuit based) may not recognize the call and will not connect it to the 911 system. Also, if a company is operating a virtual private network over your facilities and has their phone service on it, a 911 call made from that system may go to the location where the company ties to the PSTN, which may be in a different city than the caller is in.

When a subscriber gets a new phone line, the subscriber's information (from the service order) is entered into the *customer record information system* (CRIS),

and the *street address guide* (SAG). These become the *service order input* (SOI) to the 911 database. The database is stored on redundant *automatic location ID storage devices* (servers). When the customer later makes a 911 call, it is routed to a *public safety access point* (PSAP — also called a *911 tandem*) that automatically retrieves the previously stored information and displays it on screen to the operator. There are multiple PSAPs; if the nearest cannot take the call, it is automatically routed elsewhere. The call is held until released by the 911 operator — the customer cannot release the call.

6.8 DS1/E1 Transport on IP Networks

Transport of DS1 (North America and a variant in Japan) or E1 (much of the rest of the world) channels on IP networks is important due to the widespread popularity of such interfaces. DS1 is often improperly called T1, which refers to transmission of the corresponding signals only on twisted-pair wiring.

DS1 is a transmission system consisting of 24 bearer channels at a rate of 64 kb/s (one voice signal) each, plus 8 kb/s of overhead, for a total data rate of 1.544 Mb/s. The E1 standard used in Europe and many other places is based on 32 channels, of which 30 are used for data, at a data rate of 2.048 Mb/s. Though slow by modern standards, DS1 circuits are extremely popular as PBX interfaces as well as for other business data access needs. It is important for a cable TV practitioner to have a fundamental understanding of DS1 transport because it is used so frequently in business applications.

DS1 circuits were originally used to transport 24 channels of telephone calls in a time division multiplexed format, and sometimes you will see DS1 referred to as TDM. The issue in cable TV practice is frequently the transport of DS1 circuits over IP, often referred to as TDMoIP. Such transport can be tricky due to the need for tight timing in the delivery of data, which is *not* a characteristic of IP networks. By their nature, IP networks cannot guarantee the time of delivery of packets, but DS1 circuits rely on precise timing for their end-to-end performance.

6.8.1 Timing Recovery

A few methods have been developed to permit timing recovery at the user end of a DS1 network. Here we describe adaptive clocking, a method that allows DS1 circuits to be transmitted over IP networks that do not maintain precise end-to-end timing. Figure 6.19 illustrates the use of adaptive clocking in timing recovery. The 1.544-Mb/s datastream is delivered to the headend from an outside network (for example, from a telephone switch connecting to a PBX at a business). This is shown in the upper left of the figure. A clock recovery circuit supplies timing to a phase locked loop (PLL), which in turn controls a clock for the recovered upstream information, as we describe shortly.

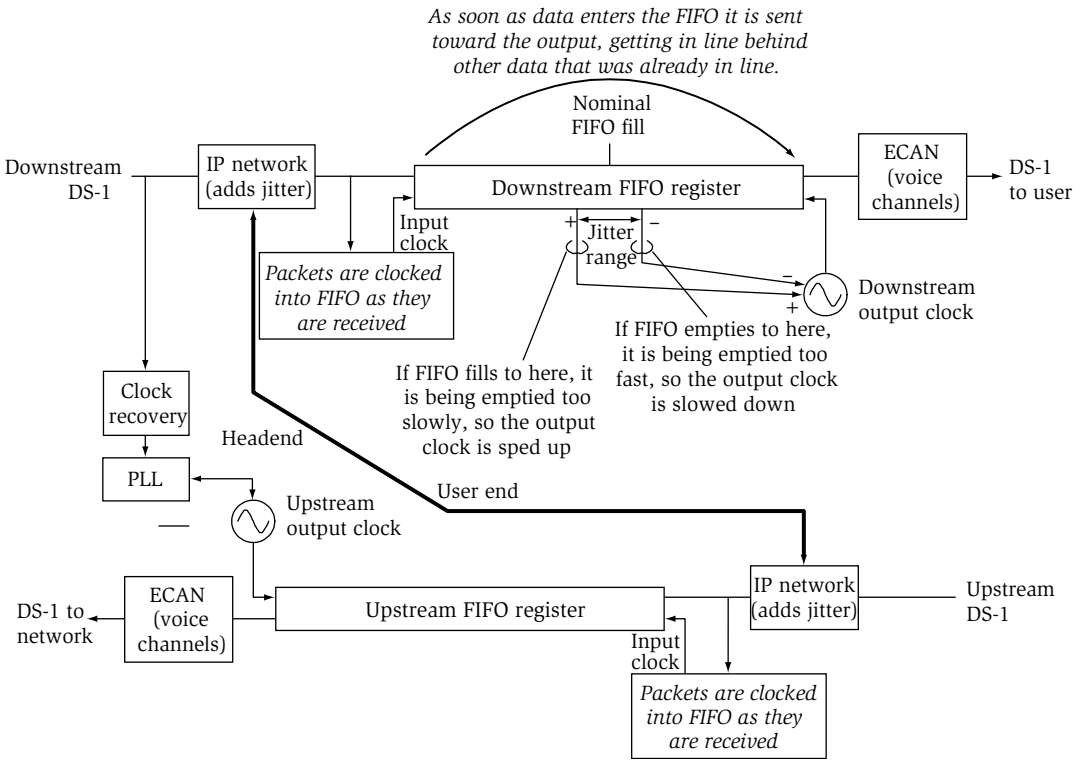

As soon as data enters the FIFO it is sent toward the output, getting in line behind other data that was already in line.
Nominal FIFO fill
Downstream DS-1
IP network (adds jitter)
Downstream FIFO register
ECAN (voice channels)
DS-1 to user
Input clock
Jitter range
Packets are clocked into FIFO as they are received
Downstream output clock
If FIFO empties to here, it is being emptied too fast, so the output clock is slowed down
If FIFO fills to here, it is being emptied too slowly, so the output clock is sped up
Clock recovery
Headend
PLL
User end
Upstream output clock
DS-1 to network
ECAN (voice channels)
Upstream FIFO register
IP network (adds jitter)
Upstream DS-1
Input clock
Packets are clocked into FIFO as they are received

Figure 6.19 Timing recovery in a TDMoIP system using adaptive clocking.

The incoming data is transmitted in the IP network, such as via a cable modem system, to a system that recovers the DS1 at the user end. At the user end, as packets arrive (not on a regular basis, but rather with jitter added by the IP network), they are clocked (loaded) into a first-in-first-out register (FIFO). This is a special-purpose shift register that has independent input and output clocks. As a packet arrives, it is clocked into the input of the FIFO, and it immediately moves as far toward the output as it can, getting in line behind packets that got there earlier. A separate output clock shifts data out of the FIFO to the using system, possibly via an ECAN, described earlier.

The output clock must run at exactly the same speed as the clock that controlled entry of the downstream DS1 data into the network, but it has no direct knowledge of that clock rate. However, the user-end recovery system can infer the clock rate by monitoring how full the FIFO is. A nominal fill point is established in the FIFO. That is, the output clock will attempt to adjust itself to the frequency that keeps the FIFO filled to the nominal fill. However, since data

arrives in packets, it is inevitable that sometimes the FIFO will fill or empty beyond this nominal fill. Thus, two thresholds are set. If the output clock is running faster than the controlling system clock, then the FIFO will empty below the "−" end of the jitter range shown. When this happens, the output clock is slowed down (that is, its frequency is reduced slightly), slowing the speed of data being shifted out. Similarly, if the output clock is running too slowly, the FIFO will begin filling more, and when it fills to the "+" limit, the output clock will speed up. Thus, over time the output clock will either speed up or slow down until it is running at the same frequency as the clock that controls delivery of the DS1 to the IP network.

If the DS1 is carrying telephone calls, it is important to provide for echo cancellation if the delay exceeds a small amount, as described earlier. This is the job of the echo canceler (the ECAN) shown at the output. Some DS1 streams carry phone calls in some time slots and other data in other time slots. The ECAN must be able to recognize what kind of data is in each time slot, and either cancel echo or not.

The downstream output clock establishes timing (which is carried in the DS1 datastream) for the user-end system. The user-end system delivers the upstream DS-1 traffic to the IP network in the lower right of Figure 6.19. At the headend, as upstream packets arrive, they are loaded into the upstream FIFO register, where they immediately move as far to the left in the upstream FIFO as possible. They are then clocked out by the upstream output clock, which is the one locked to the downstream DS-1, as shown earlier. Since the user-end clock is locked to the downstream DS-1 and the upstream output clock is locked to this same downstream DS-1, we are assured that we will not encounter a slip, in which either there is not enough data to fill the DS-1 due to late arrival or there is too much data, resulting in something having to be discarded.

It is extremely important that DS-1 circuits exhibit reliable service. If a DS-1 packet is dropped, the receiving circuit should be able to detect that fact and insert null data into the FIFO where the dropped packet should have been. This will not help with the problem of important data being dropped, but at least it will maintain timing, which is essential to the operation of DS-1 interfaces. When DS-1 packets are being carried, they must receive very high transport priority.

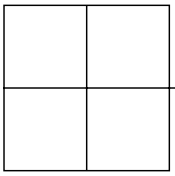
6.9 Summary

This chapter has presented a brief tutorial of the traditional telephone network, with emphasis on the interface-to-analog telephone instruments found in subscribers' homes. The chapter also presented a general description of a switched circuit cable telephony system. A fairly generic TDM/TDMA system was chosen as a frame of reference. It was the intention to describe the technology, with no

regard to one commercial system above another. Finally, we described Voice on Internet Protocol (VoIP), which is developing into the telephony method of choice. The next chapters will move into the headend, describing equipment and techniques of importance to efficient headend operation.

Endnotes

1. Nortel, The Technology of Telephony. *Telephony 101*, November 1994, Northern Telecom, Richardson, TX.
2. Much of the information in this section is taken from Whitham D. Reeve, *Subscriber Loop Signaling and Transmission Handbook*. Piscataway, NJ: IEEE Press, 1992.
3. GR-499-CORE, published by Telcordia, Morristown, NJ, issue 1, December 1995.
4. Khalid Sayood, *Introduction to Data Compression*. San Francisco: Morgan Kaufmann, 1996, p. 294.
5. J. O. Farmer and E. J. Callahan, Issues in Handling Cable Signals Within Digital Optical Interconnect Networks, *1994 NCTA Technical Papers*. NCTA, Washington, DC.
6. GR-499-CORE, published by Telcordia, Morristown, NJ, issue 1, December 1995, Chap. 10.
7. Adapted from Robert Eng, *RF Concentration: Planning/Engineering*. Customer training material, supplied by Arris Interactive, Duluth, GA.
8. Internet Engineering Task Force, RFC-2705, *Media Gateway Control Protocol (MGCP), Version 1.0*, October 1999.
9. American National Standards Institute, 11 W. 42nd St., New York, NY, 10036, or www.ansi.org.
10. Stephen A. Thomas, *IPng and the TCP/IP Protocols*. New York: Wiley, 1996, Chap. 11.
11. http://www.nena.org/PR_Publications/Devel_of_911.htm



Chapter 7
Signal Reception

7.1 Introduction

This chapter is the first of three devoted to the headend. Chapter 7 covers the ways in which signals get into the headend. Chapter 8 will cover the electronics used in processing the signals once they get into the headend, and Chapter 9 will cover topics related to the signals after they have been processed.

Chapter 7 deals primarily with antennas of various types used to capture signals for the headend, and with propagation of signals from the source to those antennas.

Although cable operators usually use antennas to receive signals, antennas are sometimes described in terms of their operation when transmitting signals. The principle of duality applies, which states that an antenna is capable of operating as either a radiator or receptor of power. This principle is based on Maxwell's equations, which describe the co-existence of radiated fields and those contained within a conductor, without defining which causes the other.

7.2 Off-Air Reception

Cable television got its start as a community antenna service, in which a large antenna was erected in a favorable place and signals from it were supplied to the community. Today, off-air channels still tend to be some of the more watched though other channels are also popular. Some cities remain prime candidates for a community antenna service since propagation conditions don't lend themselves to good off-air reception at all locations. Cities located in mountainous terrain are good examples. As cable and broadcast television have learned to cooperate, it is becoming more common for cable headends to be supplied with a direct feed from the broadcaster. That way, the so-called perils of the radio path don't take their toll on the signals. Even so, it is still common for local and distant broadcasters to be picked up using off-air antennas.

7.2.1 Antennas for Reception of Off-Air Signals

Most, but not all, off-air antennas are of one of two types: Yagis or log periodic arrays. However, before we discuss them, it is necessary to discuss the dipole, which is fundamental to the operation of many other types of antennas. The dipole will be used to introduce certain concepts that are useful throughout the chapter.

The Dipole Antenna

A dipole antenna consists of a conductor broken in the middle to allow it to be fed from a balanced transmission line. The total length of the dipole is normally one-half wavelength though other lengths are used in certain cases.

Figure 7.1 illustrates operation of a basic half-wave resonant dipole antenna. Shown in Figure 7.1(a) is a dipole antenna excited from an RF source. The balanced nature of the ideal circuit ensures that identical current will exist on each side of the dipole. Current near the ends of the dipole is of necessity small,

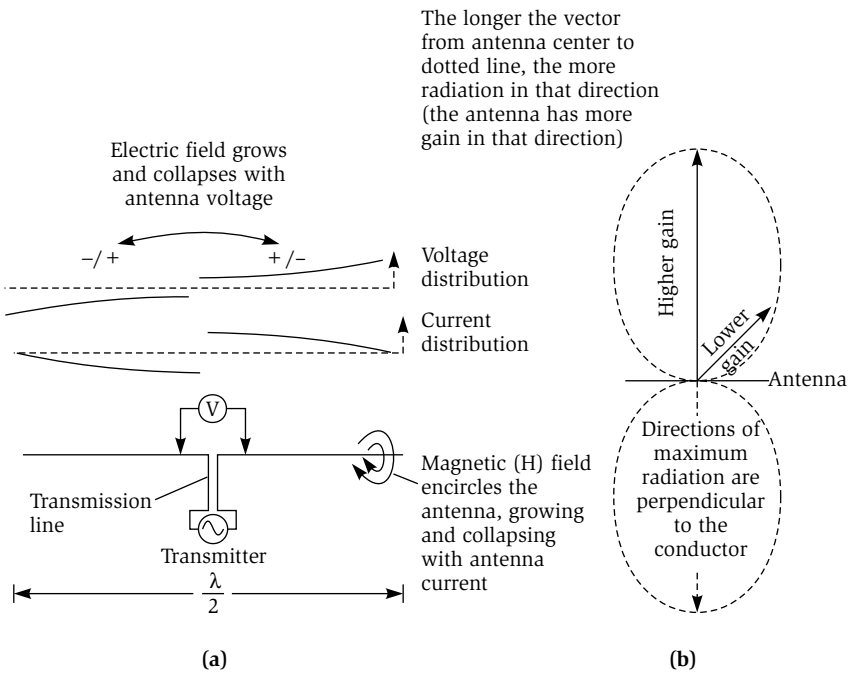

Electric field grows and collapses with antenna voltage
−/+
+/−
Voltage distribution
Current distribution
V
Transmission line
Transmitter
λ/2
Magnetic (H) field encircles the antenna, growing and collapsing with antenna current
(a)
The longer the vector from antenna center to dotted line, the more radiation in that direction (the antenna has more gain in that direction)
Higher gain
Lower gain
Antenna
Directions of maximum radiation are perpendicular to the conductor
(b)

Figure 7.1 Operation of basic dipole antenna. (a) Half-wave dipole with current distribution and oscillating doublet. (b) E plane pattern.

vanishing at the ends, because electrons, whose movement constitutes the antenna current, have nowhere to move at the very end of the wire. As you move from either end toward the middle, the RMS current magnitude increases until it reaches a maximum at the center. Similarly, the RMS voltage measured from one side of the antenna to the like position on the other side (if it could be so measured) is lowest in the middle and rises to a maximum at the two ends.

Instantaneously, the dipole is charged negatively on one side, beginning at zero and rising to a maximum charge proportional to the power supplied; then the charge decreases to zero, and that side of the antenna becomes charged positively on the next half-cycle of the exiting waveform. This process creates a rising and falling electric field from one side of the dipole to the other, which moves away from the antenna. Similarly, the current in the dipole establishes a magnetic field encircling the dipole as shown, which also moves away from the antenna. The electric and magnetic fields together form the radiated electromagnetic field.

Maxwell's equations form the fundamental mathematics describing the action of antennas and the radiation of electromagnetic energy. They say that if an alternating electric field exists, then an alternating magnetic field must exist, and vice versa. The relative magnitude of each depends on the medium in which the electromagnetic field is propagating.

Polarization

Notice that the dipole, when parallel to the surface of the earth, radiates an electric field that is horizontal. Thus, it is said that the radiated field is *horizontally polarized*. The polarization ascribed to a radiated field is always that of the electric field. It is possible to position the dipole such that the electric field is vertically polarized. If two dipoles are placed at right angles and fed with the same signal, but 90° out of phase, then the electric field (along an axis perpendicular to the plane containing the two dipoles) will rotate between vertical and horizontal polarization, and the signal is said to be *circularly polarized*. Depending on which of the two dipoles is fed with the leading phase, the circularly polarized field will rotate either clockwise or counterclockwise (viewed with the field approaching the observer), resulting in a field that is said to be either *clockwise* or *counterclockwise* circularly polarized.

Near and Far Fields

Figure 7.2 illustrates a horizontally polarized wave front in the far field. Engineers often talk about *far field* and *near field* signals. Most computations assume far field conditions. Though there is no strict definition of far and near field, in order to be considered a far field signal, the electric and magnetic fields must be in a plane at right angles to the direction of travel of the wave, and the signal must be diverging according to the square law; that is, the signal power in a

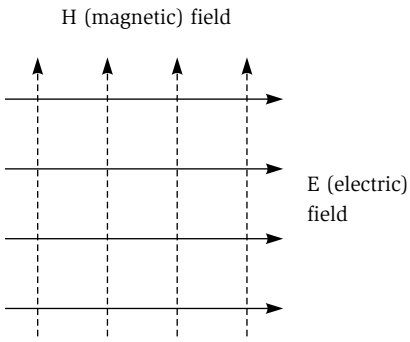


Figure 7.2 Plane wave traveling into page.

given cross-sectional area (the "power density") must be quartered every time the distance from the source is doubled. In general, these conditions are not met until the wave has traveled from the antenna by a few times the wavelength.

Note in Figure 7.1 that the E field close to the antenna is not quite flat (contained in a plane), and the H field is circling the antenna, so it cannot lie in a plane. As you go farther from the antenna, the diameter of the H field gets so large that it can be considered to lie in a plane for all practical purposes. Any magnetic field must close on itself, so the H field forms a circle. However, the dimensions of the circle become so large that for all practical purposes the H field lies in the same plane as that occupied by the E field.

Another concept of near and far field is that, in the near field, the electromagnetic field is influenced differently by different parts of the antenna, depending on the distance from the antenna part to the point of observation. That is, the antenna cannot be assumed to be at one point with respect to the point of observation. As you move far enough from the antenna, the antenna becomes more of a point source so that each part of the antenna influences the field in the same proportion as does any other part.

Dipole Pattern—E Plane

Figure 7.1(b) shows a top view of the horizontal half-wave dipole antenna, with the radiation pattern it produces. The pattern is much different if the dipole is any length other than one-half wavelength.[1] The dashed lines illustrate the relative radiation in each direction. In a direction in which the line is farther from the antenna, the antenna radiates more signal. As shown, the radiation from a half-wave dipole is strongest in a direction at right angles to the direction of the dipole. Note that Figure 7.1(b) is drawn looking down on a

horizontally polarized dipole, which has its E field parallel to the ground, and so is said to be horizontally polarized. Because the figure is drawn showing the radiation pattern such that the E field is in the plane of the pattern, this antenna pattern is said to be an *E field* pattern. Had the pattern been drawn looking at the end of the dipole, we would say that the pattern was the *H field* pattern.

Dipole Pattern—H Plane

In free space, a dipole would radiate equally in every direction when viewed end-on (the H field pattern). In the real world, dipoles are often operated close enough to ground that ground influences the pattern, and the radiation pattern would depend on how far above the ground plane the antenna was positioned. For example, a dipole located $\lambda/4$ (one-quarter wavelength) above a perfect ground has its maximum radiation at 90° (straight up). If the antenna is located at a height of $\lambda/2$, the maximum radiation is at about 30°. At a height of λ, the maximum vertical radiation angle is about 15°. Above real ground, the effect on the pattern is different, depending on ground conductivity. The vertical pattern of the antenna has to be taken into account when using a dipole to estimate the field strength of a signal.[2]

Gain of a Dipole

Since an antenna cannot amplify a signal, gain of an antenna can come only from its directionality. It must rob power from one direction of radiation in order to transmit it in another. The pattern shown in Figure 7.1 shows some directivity associated with a dipole, which means that *with respect to an isotropic radiator*, a dipole does have gain. (An isotropic radiator radiates equally in all directions.) A dipole can be shown to exhibit gain in the direction of maximum radiation, of 2.15 dBi,[3] where the suffix i indicates that the gain is stated with respect to an isotropic radiator. Antenna gain is sometimes stated with respect to isotropic, and sometimes with respect to a dipole, dBd. Thus, the gain of a dipole may be expressed as either 2.15 dBi or 0 dBd.

Practical Dipole Antennas

Figure 7.3 illustrates some practical aspects of dipole antennas. This is of interest to the cable operator because he or she is sometimes called upon to use dipole antennas to make rough field strength measurements. The length of a half-wave dipole is given by dividing the velocity of propagation *in the antenna wire*, by twice (to make the dipole one-half wavelength) the frequency of the signal to be received and/or transmitted. The velocity of propagation in the antenna wire is a function of several things, but at lower frequencies with thin wire antennas, the length may be approximated by

$$l \approx \frac{468}{f} \tag{7.1}$$

where

$$l = \text{antenna length (feet)}$$
$$f = \text{frequency (MHz)}$$

As the frequency increases, the length-to-diameter ratio drops, and the resonant length of the antenna tends to decrease compared with that given in the formula. Capacitive loading may also shorten the antenna. Since a resonant dipole in free space has an input impedance very close to 75 ohms, the return loss of the antenna can be measured with common cable equipment, and the length of the antenna adjusted until the return loss is optimized for the frequency at which the dipole is to be used. At lower frequencies, the dipole will have to be suspended from its ends, so end insulators are used to attach support cables. Depending on the support material, capacitive coupling across the end insulators may lower the resonant frequency somewhat.

In order to maintain the required balance (with respect to ground) between the two sides of the antenna, it must be fed from a balanced source (or must supply signal to a balanced load). Most equipment used in cable work is unbalanced, so it is necessary to convert from unbalanced to balanced using some sort of *balun* (*bal*ance to *un*balance). Several types of baluns are possible; the one shown is often useful at television frequencies. The coaxial cable closest

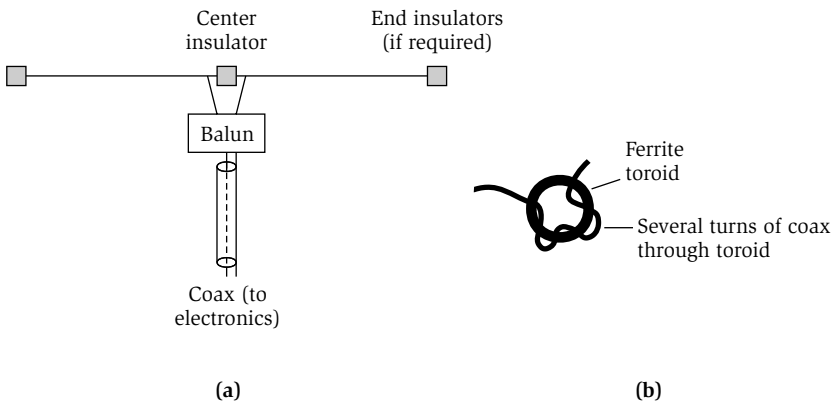


Figure 7.3 Practical dipole. (a) Practical dipole configuration. (b) One type of balun used to balance current.

to the antenna feedpoint is wrapped several times through a ferrite toroid. The resulting inductive reactance raises the impedance to any common mode current that may exist in the coax. (Common mode current is that which is in phase in the shield and center conductor of the coax.) Differential mode current, where the current in the center conductor and that in the shield are equal and out of phase, doesn't see the inductive reactance. Failure to ensure balance at the antenna driving point can cause distortion of the antenna pattern, and pickup on the coax, leading to errors in measured field strength.

A signal level meter and a half-wave dipole can be used to estimate the field strength of a radiated signal. The level of a signal is given by[4]

$$E \approx 21 f 10^{s/20} \tag{7.2}$$

where

E = field strength (μV/m)
f = frequency (MHz)
S = signal level (dBmV)

This relation is valid only for a resonant half-wave dipole in free space, oriented for maximum signal strength, and measured in a 75-ohm environment. Of course, cable and balun loss must be accounted for. Although mathematically accurate, this measurement is represented to be only an approximation to the true field strength because a number of variables can combine to reduce the accuracy. Accurate field strength measurements are difficult to make and require careful calibration of the equipment and the site. Unaccounted-for reflections alone can change the measurement by several decibels.

Even though a dipole may be used for reception of television signals, it is more common to employ a more directive antenna, which will exhibit additional gain. The gain is sometimes useful, and the directivity is almost always useful to reduce ghosting. Two types of antennas are used commonly in professional receive applications. Yagi antennas are used for single-channel reception, or at higher frequencies a Yagi may cover several channels. At lower frequencies, it is not possible to build a Yagi with enough bandwidth to cover several channels, but log periodic antennas may be used. Occasionally, other antennas will be used, but they are less common.

Yagi Antennas

The Yagi (or Yagi–Uda) antenna is named after its inventors. It is a parasitic antenna, meaning that one of several elements is driven, and current in the driven element produces current in the others. Figure 7.4 illustrates a four-element Yagi. Yagis have at least two elements, and the number can go much

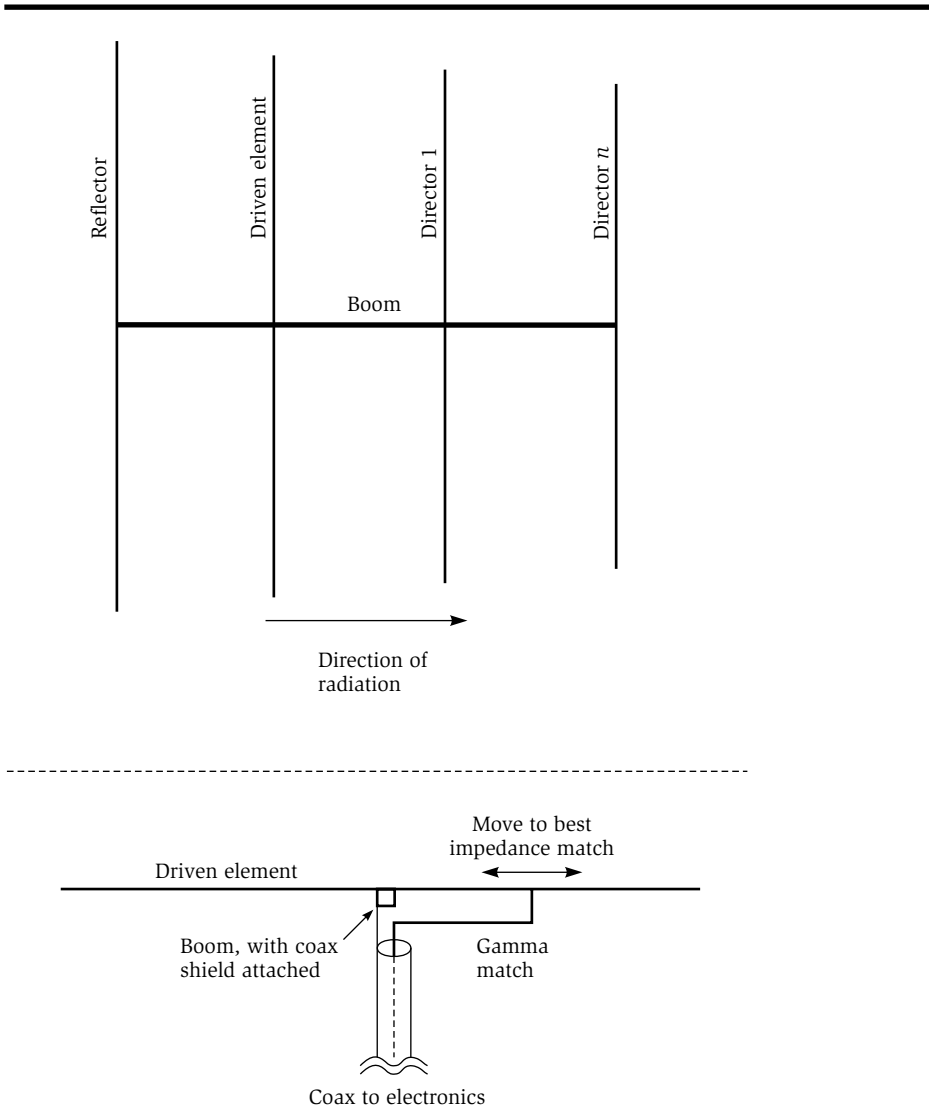


Figure 7.4 Yagi antenna.

higher, especially as the frequency increases. All else being equal, the more elements, the larger the structure, and the more directional and higher gain the antenna.

The driven element is essentially a dipole though the length is somewhat reduced owing to proximity effects of the other elements. A balun may be used

to feed a Yagi, but other feed methods are more common. Shown is a gamma match, which is used frequently, especially at lower frequencies. The coax shield is connected to the boom, which is electrically attached to the center of the driven element. The center conductor is extended along one side of the driven element to the point at which the impedance is correctly matched to that of the coax.

A two-element Yagi typically has a driven element and one reflector, located behind the driven element, where *front* is the direction of signal radiation or reception. The reflector is slightly longer than the driven element. As more elements are added, they are added in front of the driven element and are progressively shorter. A Yagi allows a number of degrees of design freedom in the spacing between elements and the length of each. Because of this, Yagi characteristics may be modified, resulting in trade-offs between gain, bandwidth, and pattern.[5]

Log Periodic Antennas

Figure 7.5[6] illustrates the other type of off-air receiving antenna commonly used in cable television, particularly at lower frequencies. This is the log periodic antenna. It can be a very broadband antenna with good pattern characteristics. The penalty is that it can get rather large. Each element of the log periodic

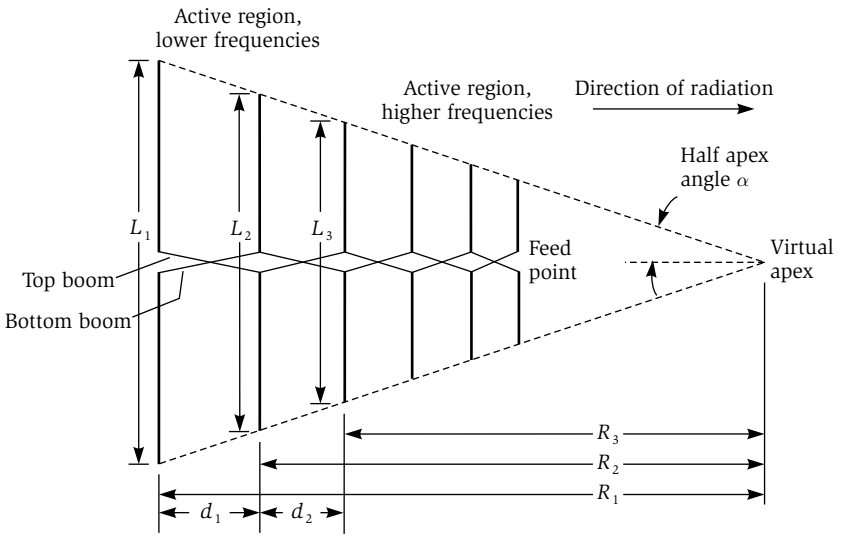


Figure 7.5 Log periodic antenna.

antenna is driven. It is driven from the front with a balanced feed. The length of each element, and the spacing between them, have a ratio relationship, τ, called the *scaling constant*:

$$L_2 = \tau L_1 \quad R_2 = \tau R_1 \quad d_2 = \tau d_1 \quad L_3 = \tau L_2 = \tau^2 L_1$$

and so on. A second constant, the spacing constant σ, further describes the antenna:

$$\sigma = \frac{d_n}{2L_n}$$

The log periodic antenna exhibits an active region for every frequency at which it operates. At lower frequencies, the active region is toward the rear of the antenna. It moves progressively forward as the frequency increases.

The log periodic antenna used for television work is usually constructed using two conducting booms, one above the other. The booms are fed at the front with a balanced signal, which propagates along the two booms to the active region. In order to feed that antenna with an unbalanced coax line, the coax is routed through one of the two booms, which then forms an infinite balun to convert from unbalanced to balanced.

Figure 7.6 illustrates a typical pattern for either a Yagi or log periodic antenna. As with all directional arrays, these antennas exhibit maximum radiation (sensitivity) in one direction. The antenna may be characterized by its *beamwidth*, the angle θ between the directions in which the power drops to one-half (-3 dB) of that in the direction of maximum radiation, or the *main beam* (sometimes called the main lobe). Ideally, the antenna would have no response in other directions, but as a practical matter, any antenna will have some response in other directions. Those responses are called *side lobes*, and an antenna will usually have several side lobes. The number of decibels that any side lobe is below the main beam is a figure of merit for the antenna. The ratio of sensitivity in the main beam to that from behind the antenna is called the front-to-back ratio. This is usually specified separately from the side lobe level. Between the side lobes, there is often an angle of minimum radiation, called a *null*.

Interference and Ghost Reduction

Directional antennas are selected for two characteristics in cable television receive applications: gain and directivity. The antennas exhibit gain, which increases the signal level at which a distant station is received, possibly improving the carrier-to-noise ratio. The directional characteristics of the antenna can be useful for reducing co-channel and ghosting interference. Co-channel interference results when a desired station is being interfered with by a more distant station operating on the same frequency. In cases of potential interference, the

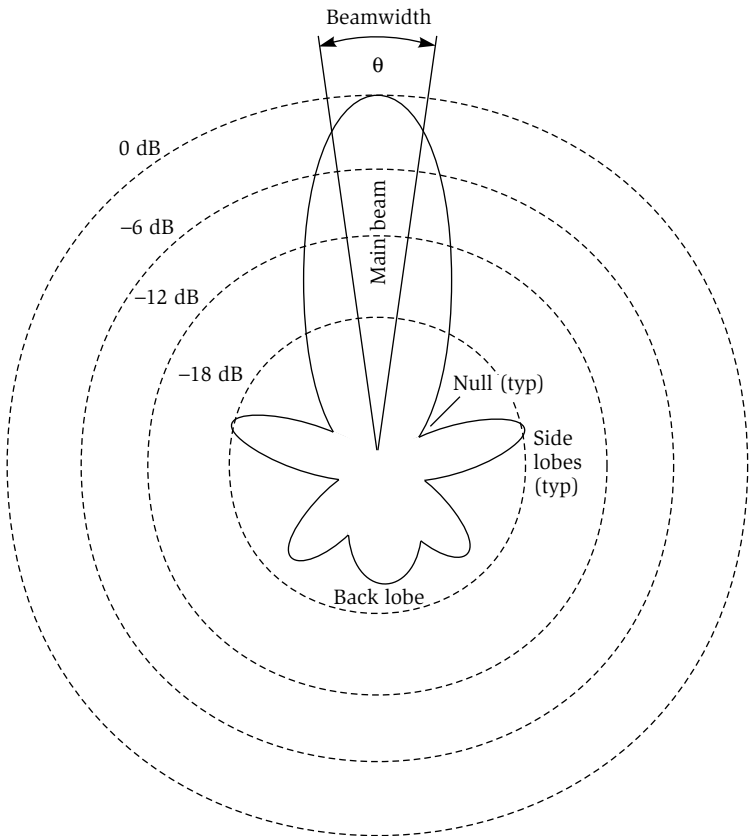


Figure 7.6 Typical antenna pattern for Yagi or log periodic antenna.

FCC assigns offsets to the stations involved, such that their picture carriers will be nominally offset from each other by either 10 or 20 kHz. It is known that such an offset will reduce the visibility of a beat. The visibility of the beat is also related to the level of the interfering signal relative to the level of the desired signal. It is often possible to reduce a co-channel beat by using an antenna with better directional characteristics so as to reject the interfering signal. Alternatively, it may be possible to rotate the receive antenna slightly so as to place the interfering signal in a null in the antenna pattern, as illustrated in Figure 7.6.

The interference may not be due to another station, but it may be due to the station being received. This can occur when a signal is reflected from some object such as a building, water tower, or mountain, and arrives at the antenna after the main signal. By the time the reflected signal arrives at the antenna, the

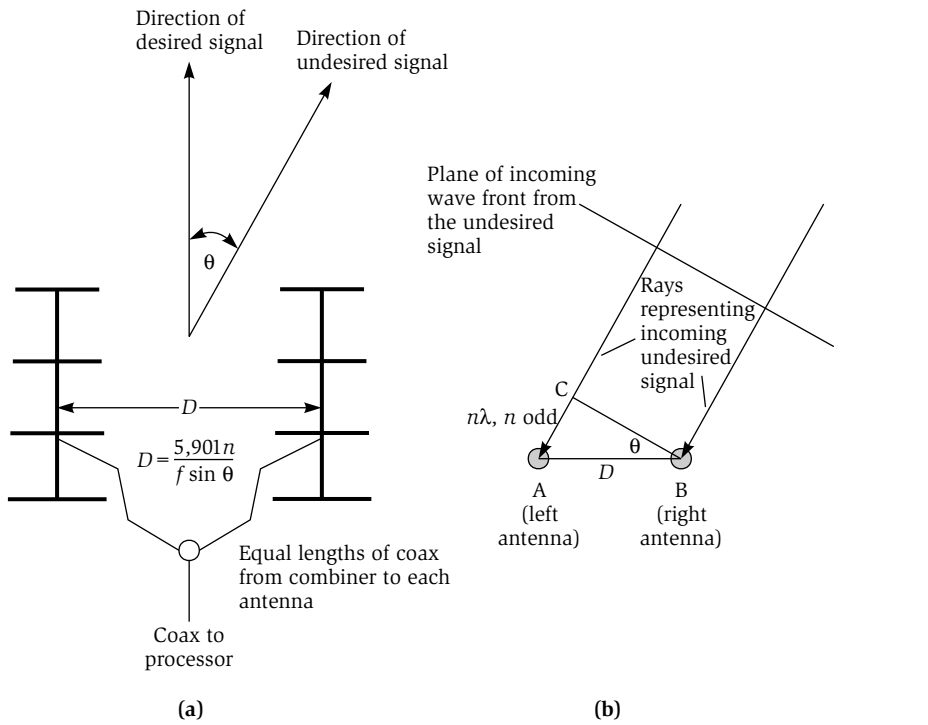


Figure 7.7 Use of multiple antennas to reject interference. (a) Antenna spacing. (b) Geometry showing why phase cancellation takes place.

electron beam in the picture tube has moved, so the reflected signal is seen as a ghost. Again, it may be possible to move the antenna slightly to put the reflected signal in a null, or it may be possible to use a more directive antenna to reject the interference.

Figure 7.7 illustrates another technique that is used to reject interference when that interference is coming from a significant angle with respect to the desired signal. Two antennas are mounted horizontally, spaced by a distance *D*, which varies with the angle of arrival of the interfering signal. As shown in Figure 7.7(a), the two antennas are fed with equal lengths of identical coax, which supply the signals to a combiner. The lengths of the two pieces of coax are not critical, so long as they are exactly the same.

Figure 7.7(b) shows how the cancellation takes place. The phase of the signal coming from the desired source is the same at the two antennas, so at the combiner, the signals from each antenna add. However, distance *D* is selected such that the signal from the undesired source will phase cancel when the signals from the two antennas are combined. This condition is satisfied when *D* is chosen

such that the undesired signal must travel one-half wavelength farther to the left antenna than to the right. The angle between the desired and undesired is θ, shown in right triangle ABC. The problem then reduces to defining D such that, given θ, side AC is an odd multiple of a wavelength at the frequency of reception (use the picture carrier frequency). This will cause the interfering signal to cancel. Substituting known relations, we can derive the formula for D:[7]

$$D = \frac{5{,}901n}{f \sin \theta} \tag{7.3}$$

where

D = distance between antennas (inches)
f = frequency (MHz)
θ = angle between desired and undesired signals
n = an odd integer

Another technique that has been used in the past to cancel an undesired signal is a bucking array: a second antenna is pointed in the direction of the undesired signal. Its received amplitude and phase are adjusted such that it cancels the undesired signal being received from the main antenna. This technique was once popular, but seems to be less so now owing to the critical need for matching the level and phase between two (usually) unlike antennas.

Yet another method that has been used for ghost cancellation requires the cooperation of the television station. The station radiates a circularly polarized signal, and the cable system receives it on a circularly polarized receiving antenna. It may be shown that upon reflection, the polarization of a circularly polarized wave changes the sense of polarization, from left hand to right hand or vice versa. A circularly polarized receiving antenna can then reject the undesired polarity.

In the 1980s, the FCC permitted television stations to transmit circularly polarized signals so as to aid in rejecting ghosting. A few stations have elected to do so, but more have not. One problem is that a new and expensive antenna is required. The antenna must be installed and fed carefully to ensure truly circular polarization. Another problem is that, if all else stays the same when a station goes to circular polarization, the radiated signal in the horizontal plane traditionally used for television transmission is halved. This means that people receiving the signal with a horizontally polarized antenna will suffer a 3-dB loss of signal level. To recover this would require another 3-dB gain in the antenna or a doubling of transmitter power. Neither is a simple or inexpensive proposition.

Offsetting these objections somewhat is the observation that a circularly polarized signal is somewhat easier to receive on rabbit ear antennas, which respond to both horizontal and vertical components. This benefit may be obtained even with elliptical polarization, which is circular polarization in which

the phase or amplitude of the vertical and horizontal components is not controlled carefully.

Yet another method that has been used to reduce reflection is the use of a ghost-canceling system. A training signal is transmitted in the vertical blanking interval of the television signal (line 19 has been designated by the FCC). This training signal is subjected to the same ghosting as is the video. A filter in the receiver is adjusted until its characteristics return the training signal to its known transmitted form. The same filter is applied to the video, removing the ghost. The system has been shown to be effective in reducing ghosts. Cooperation of the station is required to insert the ghost-canceling reference signal. This technique is simpler than going to circular polarization. Although it is being used in some places, the public has never bought a lot of television sets that have ghost-canceling circuits built in. Ghost-canceling hardware is available to the cable industry for headend use.

7.2.2 Propagation of Off-Air Signals

The best receiving antenna is of no use if signals are not present for it to receive. Signals propagate from the transmitting antenna to the receiving antenna in ways that are not totally predictable though some characteristics are sufficiently constant that we can make some statements about the propagation. Signals are reflected from natural and man-made structures so that the resulting signal at the receiving antenna comprises direct and reflected *rays*. (A ray is a path taken by a signal from one point to another.) Direct and reflected rays may add constructively (in phase) or destructively (out of phase) or in between (some arbitrary phase angle). If sufficient delay is present between the signals, a ghost is produced. If the differential delay between paths is short, a distinct ghost may not result though the video waveform will be distorted.

Free Space Propagation

In the far field, the power varies inversely with the square of the distance from the source. This may be intuitively comprehended by remembering that the power from the source is spreading over two dimensions perpendicular to the direction of propagation. If so, then the area over which the power is spread quadruples every time the distance doubles. The same power is spread over four times the area, so the power density is quartered.

Signal strength is normally measured in microvolts per meter, and the units are often expressed in decibel terms, where the unit of signal level is decibels with respect to 1 microvolt per meter (dBμV/m). A particularly useful expression describing propagation of a signal in free space is[8]

$$E = 105 + P_t + G - 20 \log d \tag{7.4}$$

where

E = field strength (dBμV)/m)
P_t = power transmitted (dB above 1 kW, dBk)
G = gain of transmitting antenna (over isotropic) in the direction of concern
d = distance from transmitter (km)

Though useful in satellite work, free space propagation is more of a reference point in discussing terrestrial propagation. Reflections from the ground and other natural obstacles, and from man-made objects, will significantly modify the equation. At lower frequencies, reflections from the ionosphere will change the observed propagation. These ionospheric effects can sometimes extend well into the VHF TV band though they tend to show the greatest effect at lower frequencies.

Effects of Antenna Height

With the assumption that ground approximates a large flat surface, an approximation to field strength as a function of antenna height is given by Hall as[9]

$$E = 2E_d \sin\left(\frac{2\pi h_1 h_2}{\lambda d}\right) \tag{7.5}$$

where

h_1 = height of transmitting antenna
h_2 = height of receiving antenna
λ = wavelength
d = distance between transmitter and receiver

To get a feel for how receiving antenna height affects received signal level, consider Figure 7.8, which plots the relative strength of a signal from Hall's equation, for channels 2 and 40 (off-air), for a 1,500-foot transmitting antenna. At channel 2, the receive antenna is not high enough, even at 100 feet, to permit destructive interference between the direct ray and one reflected from the ground. At channel 40 (off-air, not cable), you can see clearly the effects of ground reflection.

We caution you against taking this illustration literally. It assumes a flat, perfectly conducting earth with no extraneous signal reflections. None of these conditions applies, except possibly to paths across open water. It does illustrate, however, that though increasing receive antenna height can be useful, there are times when you might expect a lower signal level as the antenna height is raised.

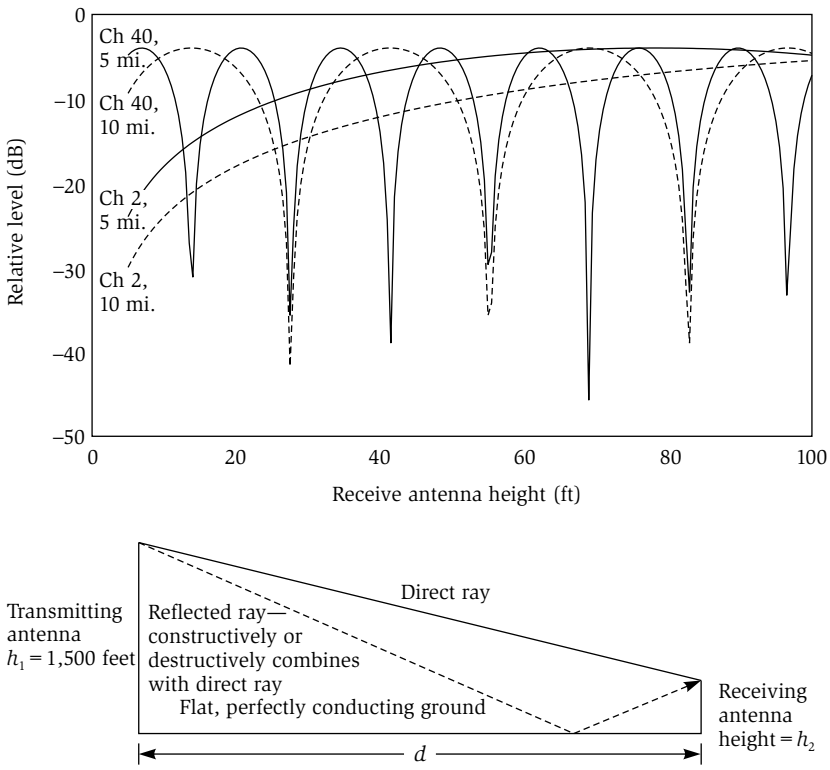


Figure 7.8 Estimated relative field strength versus receiving antenna height.

FCC Methodology for Prediction of Signal Strength

At television frequencies, radio waves are subject to shielding by the curvature of the earth though the effective radius of the earth for propagation purposes is larger than the physical radius. (See Chapter 14 for more information relating to the effective radius of the earth.) Also, one must account for reflection from the earth, which adds constructively or destructively with a direct signal. Thus, height of the transmitting and receiving antennas must be taken into account when predicting the strength of a received signal. For purposes of computing television signal strengths, the FCC uses a standard receiving antenna height of 30 feet (9 meters). The technical regulations for the broadcast industry, 47CFR section 73.699, feature graphs (Figures 9 through 11 of those rules) to be used in predicting the strength of a signal at various distances from a television station. All curves are based on a receive antenna height of 9 meters. For each frequency band, channels 2 through 6, and FM channels 7 through 13 and channels 14

through 69, two sets of curves are supplied. The first, F(50,50), is used to predict the field strength that is exceeded at 50% of potential receiver locations for at least 50% of the time. The second, F(50,10), is used to predict field strength exceeded at 50% of the potential receiver locations for at least 10% of the time. The curves are based on theory coupled with practical measurements. At higher frequencies, there may be modifications to the curves in the future.

Once a television station is on the air, it must make field strength measurements using procedures prescribed by the FCC to determine actual coverage and interference potential. These measurements are made using calibrated antennas and field strength meters not unlike the better selective voltmeters used in the cable industry. Note that a selective voltmeter by itself is not capable of measuring field strength. The old name used by the cable industry, field strength meter, is a misnomer. The selective voltmeter is not a field strength meter until it is combined with an antenna having known gain-versus-frequency characteristics, and certain computations are made (Equation (7.2)).

The FCC sets standards for the minimum field strength for two grades of service. Grade A service is generally expected to be provided to the principal community served by a television station. Grade B service is a lower field strength, but one that should result in satisfactory pictures if good outside antennas are used. Of course, further improved service may be obtained by use of better antennas and lower noise receiving systems.

To aid in planning of cable reception, computer programs are available that use the FCC's database of station coverage to predict the received signal level at any geographic coordinates. The programs are useful, but in the final analysis, the only way to know with certainty what signals exist where is to go look for them. The old cable industry axiom is still true: signals are where you find them.

7.3 Satellite Reception

Communications satellites used in the cable television service are in geosynchronous orbit, about 23,500 miles above the equator. At this altitude and above the equator, the orbital period of a satellite is 24 hours, equal to the time the earth requires to rotate one complete revolution. If a satellite is launched to this orbit, moving in the same easterly direction as the earth, it appears to stay stationary over one point on the equator. Because of slight irregularities in the earth's shape, and because insertion into an exact orbit is impossible, it is necessary to track the position of the satellite, and command small corrections as it drifts out of its assigned slot. Tracking stations operated by or for the satellite owner perform this task. Fuel stored on the satellite is consumed making these small corrections. The life of the satellite is often based on the amount of fuel it can carry at liftoff. When the fuel is used up, orbital correction becomes impossible, and the satellite gradually drifts off station, becoming useless for reception using fixed antennas.

In this section, we discuss the issues that go into installation and operation of a receive-only earth station for video. Design equations are presented that will estimate the requirements for, and performance of, an earth station. This is not intended as in-depth coverage of the material.

The two popular bands for satellite service today are C band (downlink frequencies 3,700 to 4,300 MHz) and Ku band (downlink 11.7 to 12.2 GHz). The satellites themselves consist of a number of channels called *transponders*. Each transponder is a complete subsystem so that if one fails the others are not affected. A transponder consists of an input amplifier, bandpass filter, frequency translator, and output amplifier. It is usually a single conversion device. Most satellites today have 24 transponders on board though the number may vary.

Normally, one transponder will handle one analog television signal, which is frequency modulated onto a carrier. On occasion, two signals have shared a single transponder, in what is known as a *half transponder* mode. In the more common full transponder mode, the transponder is normally operated somewhat into saturation, meaning that it is slightly clipping the signal. This doesn't bother a frequency modulated signal and ensures maximum efficiency from the transponder. Each transponder is normally 40 MHz wide, including guard band between channels, with the actual signal occupying about 36 MHz.

In order to make the best use of the frequency spectrum, frequency reuse is practiced, in which signals of two polarities are transmitted. For example, some signals can be transmitted using vertical polarization, and others can be transmitted using horizontal polarization. The polarities are said to be *orthogonal*. The dictionary definition of orthogonal is that it has to do with a right angle, or something rectangular. An antenna that is horizontally polarized is theoretically insensitive to vertically polarized signals, and vice versa.

In U.S. satellite practice, the polarizations are linear, nominally vertical and horizontal, though the axes may be rotated by the time the signal travels from satellite to earth. In many international satellite applications, the orthogonal polarizations are right- and left-hand circular polarization. Either set may be used successfully.

Satellites are assigned orbital slots by the FCC for U.S. use and by the appropriate authority in other countries. Slots are assigned to each country by common consent under the auspices of the International Telecommunications Union. The orbital slots assigned to the United States are rather crowded, and the FCC has assigned orbital slots in two-degree increments. This places burdens on antennas to discriminate against another orbital slot two degrees away from the one to which it is aimed.

7.3.1 Antennas for Reception of Satellite Signals

Many different antenna types are used for reception of satellite signals. This discussion is restricted to those commonly used in cable television work today. Variations are common between different manufacturers. Almost all antennas

are of the parabolic ("dish") type, in which a relatively large parabolic surface is used to collect signal power. That signal is focused on the actual receive antenna, known as the *feed*. At the feed, the signal is amplified by a low-noise amplifier (LNA) and usually block converted to a lower-frequency band. The device used to amplify and convert to a lower band has come to be known as a low-noise block converter (LNB). The most common band to which signals are converted is 950–1,450 MHz though other frequency plans have been used. The electronic hardware is described in more detail in Chapter 8.

The Parabola

Almost all satellite antennas are based on the parabola, whose geometry is shown in Figure 7.9. (Of course, practical antennas are realized in three dimensions rather than the two shown here for clarity. You can envision the third dimension by rotating the figure around the x axis.) The interesting characteristic of the parabola is that parallel rays approaching the parabola from the right, as illustrated, are all reflected to one point, called the *focus* or *focal point*, of the parabola. If the feed is placed at the focus, it will capture rays striking any point on the parabola, so long as those rays are coming from the direction shown. That direction is known as the *boresight* of the antenna.

The parabola as defined by the equation shown in Figure 7.9 extends to infinity along the x axis, with the slope of the parabola asymptotically approaching the horizontal. Real antennas are limited to some useful dimensions. The larger the area of the dish, the more power the antenna can collect. Antennas are characterized by their size, which is defined by the length of a line from one side of the dish to the other.

An important quality of a real dish is the degree to which it conforms to the parabolic shape. Real dishes are constructed out of some sort of metallic surface, often of pieces bolted together in the field. A figure of merit for the dish is the RMS error in the conformity to the parabolic shape. The locations of many positions on the parabola are measured and stated as an offset, in the x axis for reflector orientation as shown in Figure 7.9, from the ideal location as defined by the parabola formula. The RMS error is then computed as follows:

$$\text{error(RMS)} = \sqrt{\frac{(x_1' - x_1)^2 + (x_2' - x_2)^2 + \ldots + (x_n' - x_n)^2}{n}} \tag{7.6}$$

where

x_i' = actual position
x_i = theoretical position
n = number of points measured

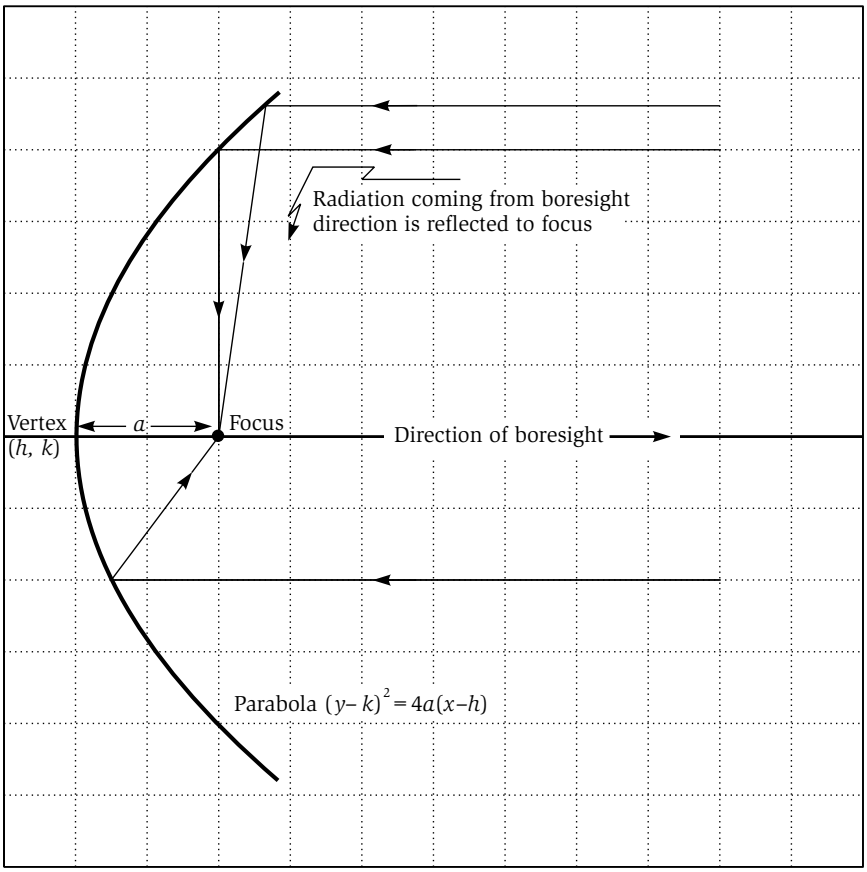


Figure 7.9 Parabolic surface.

The error may also be measured as offset positions normal to the dish surface.[10]

As the operating frequency of the dish increases, the surface errors become more important. Thus, a dish that has low enough surface tolerances to be useful at C band (3.7–4.2 GHz) may not provide satisfactory performance at Ku band (11.7–12.2 GHz). Surface errors reduce the gain of the antenna and spread out the pattern: power radiated from the feed and striking an element of the reflector not positioned correctly will be reflected in the wrong direction. The effect is distortion of the pattern. Since some of the power is going in the wrong direction, less is available in the correct direction, so the gain is lower.

Prime Focus Feed

The most basic satellite antenna type is the prime focus feed, which places the feed at the focus of the parabola, as in Figure 7.10. Support spars hold the feed in position. The feed itself and the support spars offer blockage to some of the power coming from the satellite and also reflect power coming from other directions, somewhat compromising the side lobe performance of the antenna. Nonetheless, the prime focus feed is popular until the dish size becomes quite small, where the blockage and reflections are more serious.

Note that the pattern of the feed is critical for maximizing the gain and the directivity of the antenna. The feed pattern should ideally illuminate the entire surface area of the dish evenly, as shown in the dotted pattern. If the feed pattern is too narrow, the outer portions of the dish are not illuminated, so they cannot

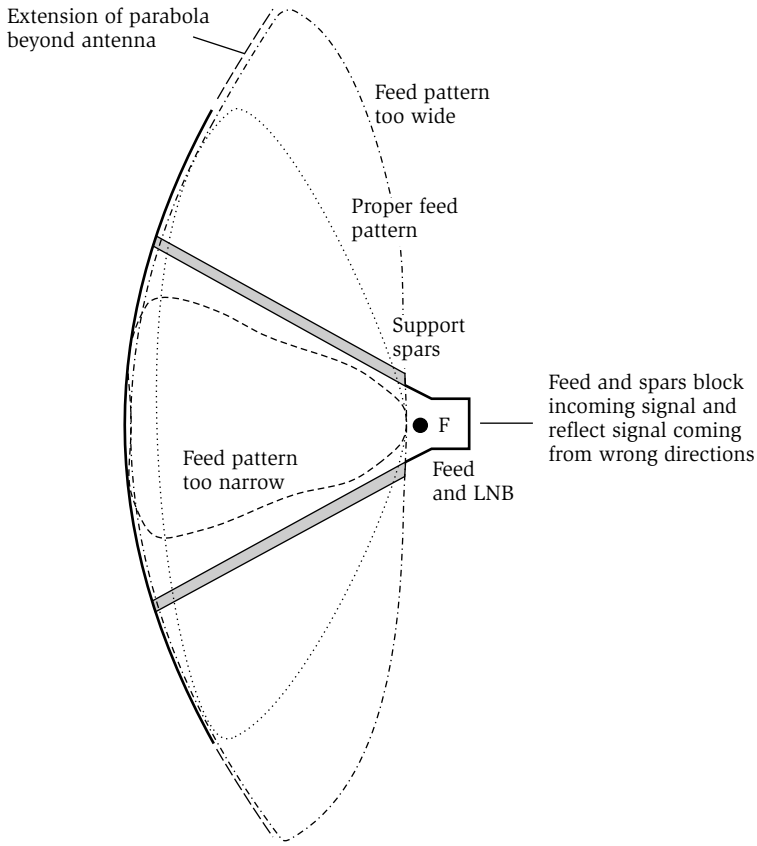

Extension of parabola beyond antenna
Feed pattern too wide
Proper feed pattern
Support spars
Feed and spars block incoming signal and reflect signal coming from wrong directions
F
Feed pattern too narrow
Feed and LNB

Figure 7.10 Prime focus feed.

aid in focusing a transmitted signal. Similarly, a received signal will not be captured if it strikes the dish outside the feed pattern because the power will be rejected by the feed when it reaches the focal point. If the feed pattern is too wide, some of the power radiated from a transmitting antenna will miss the dish surface and go off in the wrong direction, spoiling the pattern of the antenna. Received power will not be collected efficiently because the feed antenna gain is lower than optimum in the direction of the dish surface.

Matching the feed pattern to the dish is an art in itself. Of course, no feed is perfect, but the closer you can come to the ideal, the better the antenna gain and pattern. Feed antennas include dipoles above ground planes, Yagi or log periodic structures, and microwave feed horns.

Prime focus feeds are relatively simple to build and align. A disadvantage is that the feed, which contains the active electronics of the LNB, is relatively inaccessible, being held several feet in front of a parabolic surface that is facing up at some angle. Some would argue that the feed in this location is subject to relatively more damage from electromagnetic pulses that occur when lightning strikes.

Cassegrain Reflector

Another very common type of antenna is the Cassegrain reflector derived from the telescope design by William Cassegrain. It is the same basic antenna as the prime focus feed. The difference is that a second reflector is placed between the vertex and the focus of the parabolic reflector. The most common type of Cassegrain system is illustrated in Figure 7.11. A hyperbolic subreflector is used to reflect signals from the main reflector to the feed. The feed is located either at or in front of the vertex, depending on the design of the system. The hyperbolic subreflector is located such that its first focus coincides with F1, the focus of the parabolic main reflector. The second focus of the hyperbola is the location of the feed. (The second focus is from the solution of the hyperbolic equation defining the subreflector.)

The Cassegrain reflector places the feed electronics (usually a low-noise block converter) in a position where it is considered to be more protected and easier to access. With a Cassegrain configuration, it is not necessary to route the signals along a spar, leading to somewhat lower cable loss. On the other hand, this is not especially critical with the block conversion systems commonly used in cable television. The subreflector must be supported by spars, and the subreflector and spars create blockage and pattern degradation, as in the case of the prime focus feed. The Cassegrain system is somewhat more difficult to align since the installer must align not only the feed but the subreflector as well.

In general, you would choose a prime focus feed where low side lobes are important and a Cassegrain where gain for a given size antenna is most important. This latter statement is true for antennas whose diameters are large compared with the wavelength of the signal being received (on the order of 40 wavelengths); smaller antennas usually suffer too much diffraction on the edges

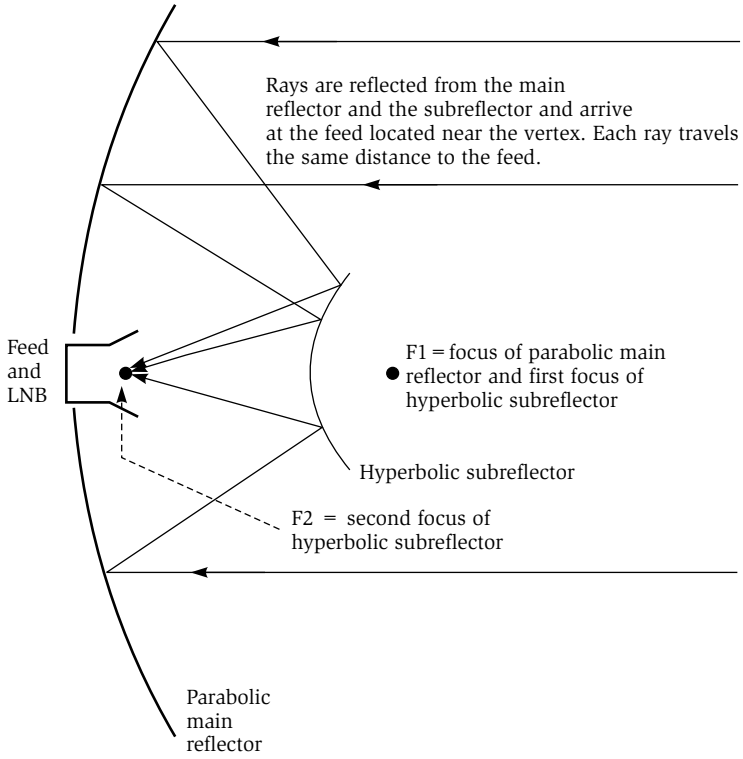


Figure 7.11 Cassegrain reflector.

of the subreflector to make the Cassegrain antenna useful. Gain is enhanced in Cassegrain antennas because the phase of a plane wave impinging on the antenna is preserved: each ray of the signal travels the same distance to the feed. This is not true of prime focus systems, but their generally smaller blockage of the primary reflector results in a better pattern.[11]

Offset Feeds

As the size of the dish gets smaller, the blockage and reflection of the spars and feed or subreflector become more of a problem. The solution can be an offset feed. The offset feed is simply the top portion of the prime focus antenna, with the bottom portion cut off. Figure 7.12 illustrates an offset feed. The illustrated system uses exactly the top half of the parabola though some offset feeds make use of less of the parabola.

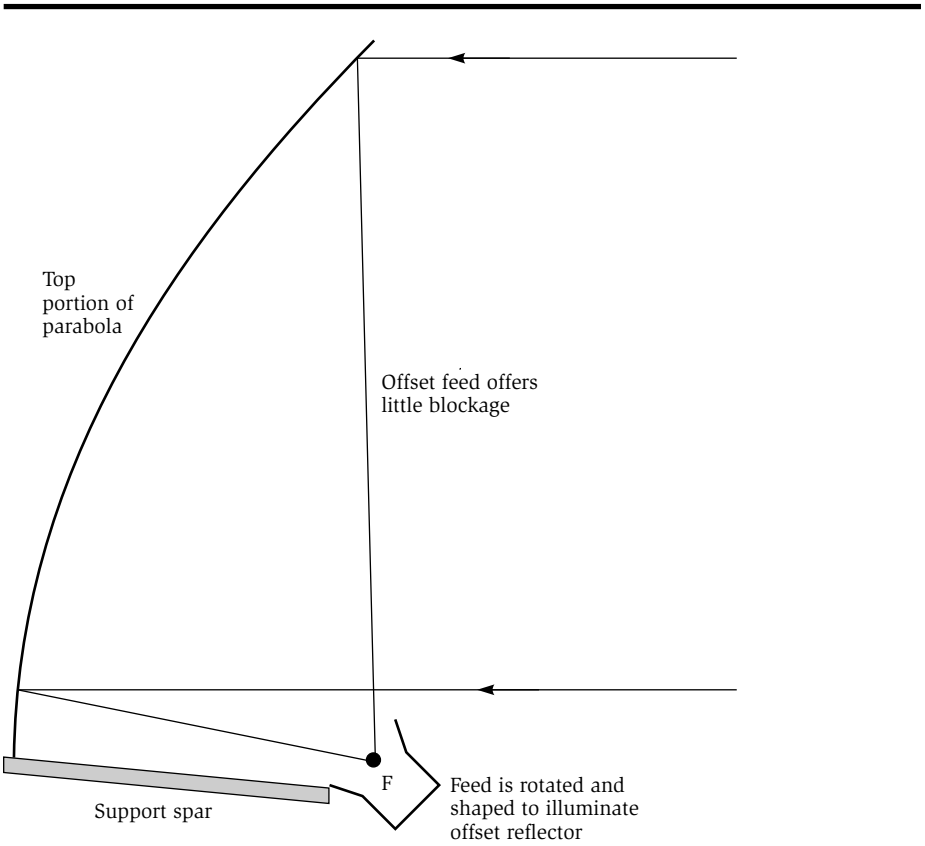


Figure 7.12 Offset feed.

The offset feed tends to be somewhat difficult to construct and align but in smaller dish sizes is a good choice because of the lack of blockage. There is some reflective pattern distortion from the spar and feed structure, but not as much as with symmetrical antennas. Because of the lack of blockage, this feed type is popular in very small aperture antennas, or very small aperture terminals (VSATs), and direct broadcast satellite (DBS) antennas.

Other Antenna Types

Several other variations of satellite antennas are used. Multiple feeds may be mounted on one reflector, which can be either a prime focus parabola or one of several shape variations. Each feed is able to look at a different satellite, so long as the satellites are close to each other in orbit. Compared with a single feed system of similar size, gain and side lobe performance are sacrificed, but the

technique can be used to reduce costs and the amount of land needed for multiple dishes. A particularly interesting multiple feed antenna is a torus antenna. The reflector has a circular contour in the scan plane and a parabolic contour in the orthogonal plane. The scan plane must be carefully aligned with the orbital arc plane. However, once aligned, that antenna is capable of quite good performance over a wide section of the geosynchronous arc.

Shaped reflectors may be used rather than parabolas or hyperbolas to optimize performance of the antenna in some respect. Of course, the feed must be designed to properly illuminate whatever shape is used.

Shrouds are sometimes placed around the circumference of an antenna to reduce the magnitude of side lobes. These antennas are sometimes called ultra directive antennas (UDAs).

Horn antennas are sometimes used. They tend to be big and difficult to align, and require awkward supporting structures. However, their performance in terms of gain and side lobe suppression can be impressive. Horn antennas are used in certain high-density terrestrial microwave applications. They look something like gigantic cones with a dome on the wide end. A feed in the apex of the cone launches a signal, which is reflected from a shaped reflector at the wide end.

7.3.2 Earth Station Geometry

In this section, we deal with the computation of pointing angles from an earth station to a satellite, and the signal loss in the downlink. Following this, we present computations to illustrate the performance you can expect from an earth station.

Many different antenna mounts are in use. Each has advantages and disadvantages, proponents and opponents. The description here concentrates on the elevation over azimuth coordinate set, a traditional coordinate set that is relatively easy to understand and work with. Antennas designed for other than communications with synchronous satellites often employ this set of axes. Antennas designed for communications with synchronous satellites may use different axes to permit the antenna to be moved between satellites by changing only one adjustment. Such systems tend to have proprietary elements, so you should consult the manufacturer for instructions on aiming the antenna.

The elevation-azimuth (or "el over az") coordinate system described here has its lower plane of rotation (azimuth) parallel to the local horizontal plane. A weather radar antenna, for example, rotates on the azimuth plane. The second axis, the elevation axis, is at right angles to the azimuth axis so that the antenna tilts up and down on the elevation axis.

Three pieces of information are of interest in defining how to point to the satellite: the azimuth (A) of the antenna, the elevation (E) of the antenna, and the path distance (D) to the satellite (also known as the *slant range*). These may be computed from knowledge of the satellite longitude (remember that the satellite latitude is zero, that being the only latitude for geosynchronous orbits). You must also know the earth station longitude and latitude.

Let the following definitions apply:

X = site latitude (degrees)
Y = site longitude (degrees)
Z = satellite longitude (degrees)
A = earth station azimuth angle (degrees)
E = earth station elevation angle (degrees)
D = slant range from site to satellite
C = difference between satellite and site longitudes

First compute the difference in longitude between the satellite and the earth station (the "site"):

$$C = Z - Y \tag{7.7}$$

Then the earth station azimuth angle is given by

$$A = 180^\circ + \tan^{-1}\left(\frac{\tan C}{\sin X}\right) \tag{7.8}$$

The elevation, E, of the antenna is given by

$$E = \tan^{-1}\left(\frac{\cos C \cos X - 0.15126}{\sqrt{\sin^2 C + \cos^2 C \sin^2 X}}\right) \tag{7.9}$$

By convention, longitude is positive west of the prime meridian and negative east. Latitude is positive north of the equator and negative south. The earth station azimuth angle is measured from true north, rotating clockwise as viewed from above the earth station. Use true north, not magnetic north, as the reference. The elevation angle is zero if the antenna is pointing horizontally and 90° if it is pointing vertically.[12] The constant 0.15126 is the ratio of the earth's radius to the satellite's orbital radius.

The slant range, or distance from the site to the satellite, is given by[13]

$$D = 26{,}199\frac{\cos C \cos X - 0.15126}{\sin E} \text{ statute miles}$$

or

$$D = 42{,}164\frac{\cos C \cos X - 0.15126}{\sin E} \text{ kilometers} \tag{7.10}$$

7.3.3 Earth Station Performance Computations

From the slant range and certain other information, we can compute the expected performance of an earth station. The signal level to be expected at the antenna may be computed from the published effective isotropic radiated power (EIRP) of the satellite and the path loss, for which the slant range is needed.

Concept of EIRP

EIRP is the power that would have to be supplied to the antenna *on the satellite* if that antenna radiated equally in all directions. If the antenna on the satellite radiated equally in all directions, a typical transmitter output power on the satellite would be somewhat above 3,000 watts. This is hardly practical, but then it is not necessary for the satellite to radiate signal in all directions. It is necessary only to radiate signal in the direction of that portion of the earth that is to receive the signal. By concentrating power in that direction, the power required on the satellite is reduced drastically.

For each satellite, the owner publishes a contour map showing the EIRP at locations within the satellite's *footprint*, or intended area of coverage. This map is consulted to determine the EIRP at the location of the receive station.

Path Loss

The next quantity required is the path loss, sometimes called spreading loss because it is caused by the signal from the satellite spreading as it travels through space. Path loss may be computed from

$$L = 20 \log\left(\frac{4\pi Df}{c}\right) \text{dB} \tag{7.11}$$

where

$$\begin{aligned} D &= \text{slant range (from Equation (7.10))} \\ f &= \text{frequency (Hz)} \\ c &= \text{speed of light} \end{aligned}$$

Slant range and speed of light must be in the same distance units. From the equation, path loss increases with increasing frequency. However, gain of the receiving antenna increases as frequency increases, keeping the required antenna size the same. If the slant range is expressed in miles and the frequency in GHz, Equation (7.11) reduces to the form shown in Equation (14.12):

$$\text{path loss(dB)} = 96.6 + 20 \log(f) + 20 \log(D)$$

where

$$f = \text{frequency (GHz)}$$
$$D = \text{distance (statute miles)}$$

Station G/T

The final quantity that must be known is the G/T of the earth station. G/T (pronounced "G over T") expresses the gain of the station, divided by its noise temperature. Noise temperature is a particularly useful expression of the noise contributed by the earth station from all sources. It is not a physical temperature, but rather is the temperature that will produce the same equivalent noise in a resistor. If a resistor is connected to a power meter, the power meter will read a certain power, given by

$$P_n = kTB \tag{7.12}$$

where

$$P_n = \text{noise (watts)}$$
$$k = \text{Boltzmann's constant} = 1.38 \times 10^{-23} \frac{\text{Joules}}{^\circ\text{K}}$$
$$T = \text{temperature } (^\circ\text{K})$$
$$B = \text{bandwidth (Hz)}$$

This noise power, P_n, is known as thermal noise, and is a basic law of physics: a resistor at a certain temperature will produce a certain amount of noise according to this equation. This doesn't have to be a physical resistor: it is often the terminals of an antenna, which looks electrically like a resistor. Real circuits, such as LNBs, add additional noise above and beyond the noise that would be predicted solely from the thermal noise. Noise temperature is a measure of the amount of excess noise, over and above thermal noise.

Other sources contribute to the noise temperature of a receive station. Space noise, which is produced by a number of objects in space, is a factor, but with small earth stations, it is usually a small factor. An exception is the sun, which is a very significant noise-producing object in space: when the receiving station, satellite, and sun are lined up, communication from the satellite is lost, overpowered by noise from the sun. Also, the earth has a rather high noise temperature. When the elevation of the antenna is less than perhaps 20° or 30°, the earth's noise temperature can be significant. Of course, the contribution from the earth depends on the antenna pattern in the direction of the earth.

In addition, the LNB produces excess noise over and above what you would expect from thermal noise. The excess noise can be translated to an equivalent noise temperature by solving Equation (7.12) for T and substituting the equivalent excess noise at the input of the LNB for P_n. An equivalent concept some what more familiar to cable television engineers is that of noise figure. Noise figure expresses the same thing that noise temperature expresses, but in satellite work, we often work with noise temperature because certain computations are simplified.

The G/T of a station is a function of the gain of the dish, the pattern of the dish if the elevation angle is low, the noise figure of the LNB, and any losses in the structure. Technically, losses after the LNB must also be taken into account. However, they tend to have only secondary effects on the G/T of the station unless the loss between the LNB and the receiver begins to approach the gain of the LNB. The manufacturer of the station can supply appropriate G/T values.

Higher G/T will provide better carrier-to-noise ratio, and hence better baseband signal-to-noise ratio on the video. G/T can be increased by going to a larger dish size, improving the efficiency of the dish, or using an LNB with a lower noise temperature. Some slight improvement may be obtained by reducing losses following the LNB.

Carrier-to-Noise Ratio of Received Signal

Combining the preceding concepts, the carrier-to-noise ratio of the receive station can be expressed as

$$\frac{C}{N_0} = EIRP - L + \frac{G}{T} + 228.6 \tag{7.13}$$

where

$$\begin{aligned} \frac{C}{N_0} &= \text{carrier-to-noise ratio (dB) per hertz} \\ EIRP &= \text{satellite effective isotropic radiated power (dBW)} \\ L &= \text{path loss (dB) from Equation (7.11)} \\ \frac{G}{T} &= \text{figure of merit for earth station (dB/°K)} \end{aligned}$$

All quantities are expressed in decibels. The carrier-to-noise ratio expressed is per hertz. As shown in Equation (7.12) for noise power, the total noise is proportional to the bandwidth over which the measurement is made. The measurement bandwidth is the *effective noise bandwidth* of the receiver. Noise bandwidth is that bandwidth a filter would have if it were perfectly rectangular and admitted the same noise as does the real filter.

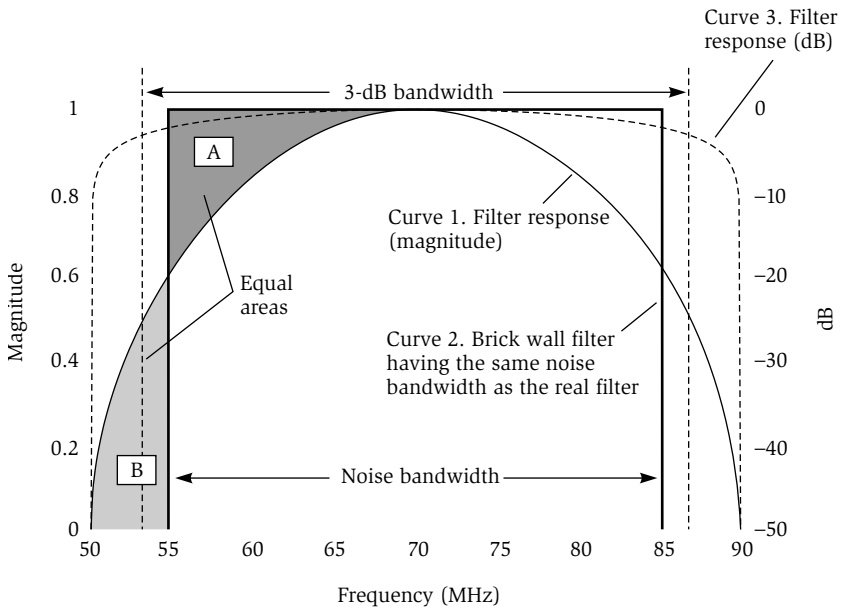


Figure 7.13 Filter response and noise bandwidth.

Figure 7.13 illustrates the concept of noise bandwidth and how it relates to the actual bandwidth of a filter. The majority of the filtering comes from the IF filter in the earth station receiver. A filter response is shown expressed in decibels in dashed curve 3. The same filter response expressed as a magnitude ratio, [(power out)/(power in)], is shown as a thin solid line in curve 1. In determining the noise bandwidth of the filter, we work with the ratio quantity, not the decibel quantity. Filter bandwidth is usually expressed at the −3-dB points, corresponding to the half-power (10 log 0.5) response. The −3-dB bandwidth of this filter is 33.6 MHz. On the other hand, the noise bandwidth is a little narrower, 30.5 MHz. The noise bandwidth is computed by calculating the area under the magnitude response curve, then determining the bandwidth of an ideal "brick wall" filter that has the same area. Curve 2 is that of the ideal filter enclosing the same area as curve 1.

In order for curves 1 and 2 to enclose the same area, area A must equal area B. Area A defines the region where the actual response, curve 1, is below the brick wall response, curve 2. Area B defines the opposite condition, where the actual response exceeds that of the brick wall response.

Note that the noise bandwidth is an artificial quantity not representative of the actual filter response. It is merely a mathematical construct used to permit computation of the noise in the bandwidth of the receiver. This quantity is needed to compute carrier-to-noise ratio from the result of Equation (7.13).

Equation (7.13) is expressed in terms of carrier-to-noise ratio in a 1-Hz bandwidth. From Equation (7.12), the total noise is proportional to the bandwidth of the system, which is expressed by the effective noise bandwidth of the filter as shown. In order to correct Equation (7.13) for the noise bandwidth, a term must be added for the noise bandwidth. The manufacturer of the receiver will supply the noise bandwidth of the receiver. Equation (7.13) then becomes

$$\frac{C}{N} = EIRP - L + \frac{G}{T} + 228.6 - 10\log BW_n \tag{7.14}$$

where

$\frac{C}{N}$ = carrier-to-noise ratio seen by receiver
$EIRP$ = satellite effective isotropic radiated power (dBW)
L = path loss (dB) from Equation (7.11)
$\frac{G}{T}$ = figure of merit for earth station (dB/°K)
BW_n = noise bandwidth (Hz)

FM Receiver Characteristics

Now consider the satellite receiver. The characteristic of interest is the baseband signal-to-noise ratio versus the RF carrier-to-noise ratio. By convention, baseband signals are characterized by a *signal*-to-noise ratio (S/N), whereas RF (and IF) signals are characterized by a *carrier*-to-noise ratio (C/N). At this point, we need to determine the receiver output S/N given the C/N computed in Equation (7.14).

Figure 7.14 illustrates a typical curve that relates baseband signal-to-noise ratio to IF carrier-to-noise ratio.[14] The measuring conditions must be stated and usually follow closely the transmission characteristics nominally used in satellite transmission (recommended by the CCIR):

Reference deviation: ΔF peak = 10.75 MHz
S/N per CCIR 421–3
525-line preemphasis
IF noise bandwidth: according to receiver (32.5 MHz in this example)

The solid line in Figure 7.14 represents the actual signal-to-noise ratio measured on the test receiver. At higher C/N, it is a straight line. As the C/N decreases, the S/N departs from the straight line, falling off more rapidly. If we

were to extrapolate the curve back to a C/N of 0 dB (dotted line), we would find that the S/N was strangely high for a C/N of 0. The point at which the actual curve diverges from the straight line is called the *threshold* of the receiver. In order to provide a repeatable measuring point, the threshold is defined as the point at which the two curves depart by 1 dB.

In determining the needed characteristics of an earth station, it is necessary to establish a nominal operating point sufficiently above threshold so that we can tolerate real-world reductions in performance. Pointing errors can cause signal loss. Polarization discrimination can suffer because of Faraday rotation and rain, causing loss of discrimination against the cross-polarized signals. Signal can be attenuated because of rainfall: the more rain the signal travels through and the harder the rainfall, the greater the attenuation.

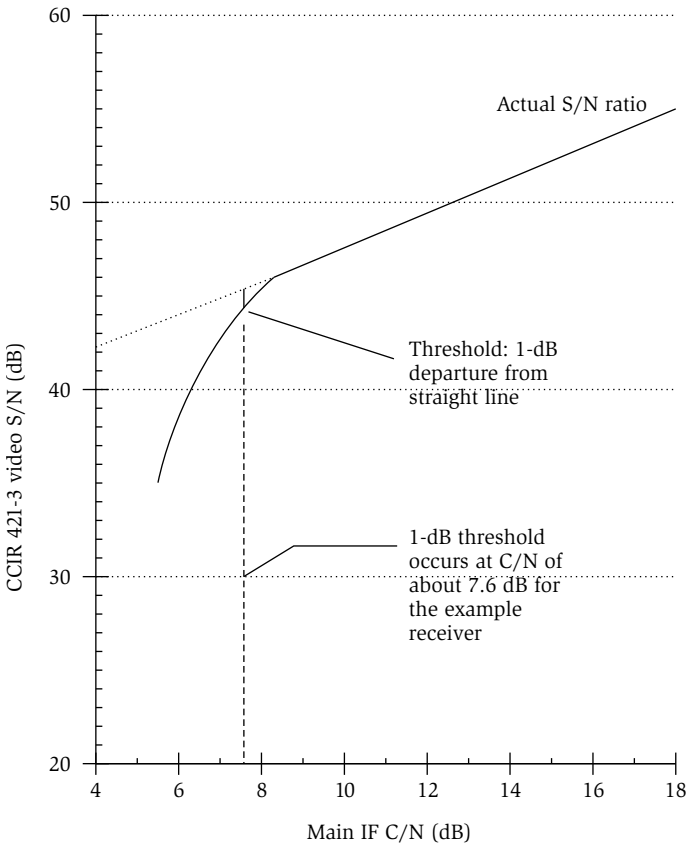


Figure 7.14 Typical receiver S/N versus C/N.

Terrestrial interference is a way of life at C band though arguably it may not be as severe as it was before many applications began migrating to fiber-optic cable. The C band is shared with terrestrial communications. Terrestrial interference has been known to make some receiving sites unusable on some transponders. However, in most cases, it can be controlled by erection of shielding fences or by putting the dish in a depression in the earth. If the interference is not too close to the center of the channel (the carrier), then sometimes a trap inserted in the IF of the receiver may prove effective at mitigating the interference. Many receivers include an IF loop-out feature to allow insertion of such traps.

Threshold Extension

In order to make use of smaller and smaller earth stations, which have lower and lower G/T, it is useful to find techniques that will enable the threshold to be reduced, enabling operation at lower C/N. Several proprietary techniques exist to provide a receiver with *threshold extension*. Regenerative techniques have been used, but a common threshold extension technique involves narrowing the noise bandwidth of the IF filter in the receiver. Some manufacturers have used a circuit that steers the filter, based on the luminance content of the scene. You can narrow only so far before the picture is adversely affected.

A common symptom of too low an IF bandwidth is streaking of colors in brightly colored areas of a scene. Red areas of the picture are particularly susceptible. This is a symptom of picture power that falls outside the passband of the IF filter of the receiver. If the filter was effectively widened, this effect would be reduced, but a higher carrier-to-noise ratio (higher station G/T) would be required.

Similarly, black dots ("sparklies") in the picture can indicate too narrow a filter bandwidth: as the carrier deviates with picture content, some power falls outside the filter passband, or the filter response is severely reduced at the frequency of the power. The effect is very similar to that observed when a station is operated too close to the threshold.

Pre- and Deemphasis

Frequency modulated signals have the characteristic that, when demodulated, the noise spectrum is triangular, increasing at frequencies farther removed from the carrier. In order to overcome this, many FM systems include preemphasis of the signal before modulation and deemphasis after demodulation. In audio work in North America, the traditional preemphasis network is a single break having a time constant of 75 μs. In satellite video work, the International Telecommunications Union prescribes a preemphasis curve that breaks upward at about 150 kHz, and breaks toward a flat response at about 860 kHz.

Figure 7.15 illustrates the recommended preemphasis curve. This shows that the color subcarrier is amplified nearly 15 dB with respect to low-frequency luminance information. Recall that this emphasis applies only between the

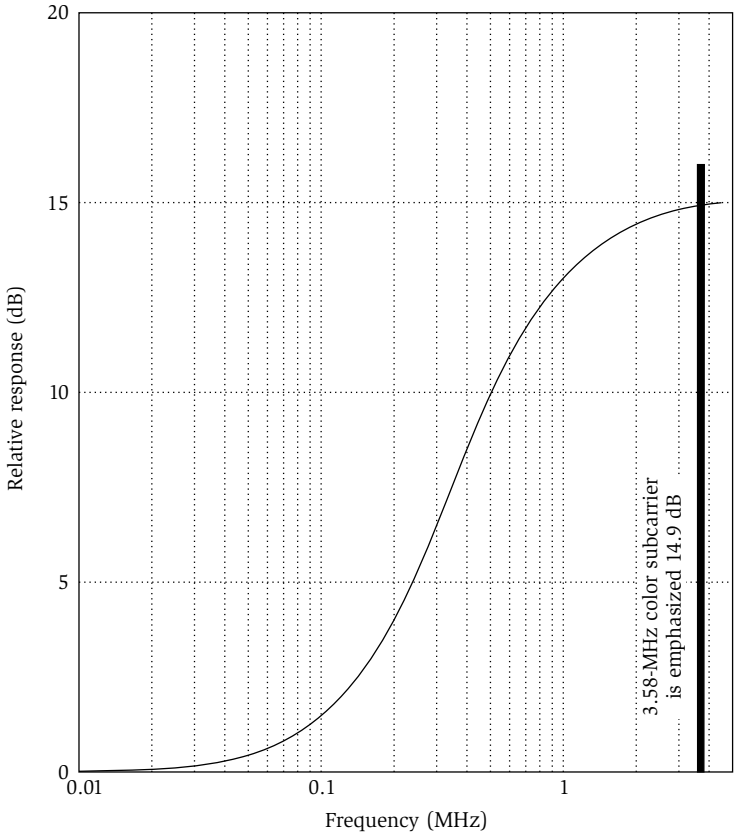


Figure 7.15 CCIR recommendation 405-1 preemphasis curve for 525-line systems.

uplink exciter and the receiver at the headend. It does not apply to the signal out of the receiver. A deemphasis circuit in the receiver restores the signal to its flat frequency response.

However, this does illustrate that the majority of the frequency deviation is taken up by the color subcarrier in the satellite portion of the transmission path. In computing the relationship between RF carrier-to-noise ratio and baseband signal-to-noise ratio, the deemphasis must be taken into account. This is done in Figure 7.14.

Audio Transmission

In the early days of satellite transmission, signals were sent in the clear. To prevent theft of signals, it became necessary to scramble the signal. The scrambling system currently employed replaces the horizontal sync with digitized,

encrypted audio. Enough bandwidth is available to provide for stereophonic transmission. If additional audio channels are required, as for cue tone and second language transmission, subcarriers are used. The subcarriers were used for transmission of the primary audio prior to the advent of scrambling.

A number of frequencies have been used for sound subcarriers, but the most common in North America are 6.8 and 6.2 MHz. The audio source is modulated onto the subcarrier (normally using the standard 75-μs preemphasis), then the subcarrier is combined with the video to form the complete signal transmitted through the satellite.

Digital Satellite Reception

As shown in Chapter 5, digital satellite transmission is primarily QPSK, with some movement toward 8-PSK transmission. The satellite transponder is simply a frequency translator and power amplifier. As such, it can handle analog or digital signals. However, in order to achieve maximum efficiency when digital signals are transmitted, the transponder is operated in or close to compression, the input signal level above which an increase in input signal does not result in a similar increase in output level. This means that we have to use modulation formats that don't depend on amplitude changes to communicate information. 8-PSK and QPSK meet this requirement.

Because the satellite transmission is digital, there is no color carrier and no audio carrier, this information being carried in the digital datastream with the video and supplemental information. Threshold extension is not practiced, and there is no preemphasis and deemphasis.

You can plot graphs similar to that in Figure 7.14 for digital transmission, where the operational parameter on the vertical axis is E_b/N_0, explained in Chapter 4. E_b/N_0 plays the same role as does carrier-to-noise ratio, normalized for the data rate transmitted. The E_b/N_0 required depends on the degree of error correction being used. Table 7.1[15] shows the expected performance of a

Table 7.1 Expected E_b/N_0 for Various Error-correction Rates

Inner code rate	*Required E_b/N_0 (dB), BER $= 2 \times 10^{-4}$ after Viterbi, QEF, after Reed–Solomon*
1/2	4.5
2/3	5.0
3/4	5.5
5/6	6.0
7/8	6.4

demodulator for different error-correction levels. A safety margin of about 2 dB on top of these E_b/N_0 values is recommended.

7.3.4 Satellite Frequency Plans and Interference Issues

The more established band for satellite reception is the C band, 3.7 to 4.2 GHz. This band is shared with terrestrial microwave links, and so interference becomes a significant issue. The issues are probably not as severe as they were a few years ago, however, since more terrestrial services have been migrating to fiber-optic cable.

Polarization Diversity and Frequency Plans

Figure 7.16(a) illustrates the frequency plan used at C band, effectively mandated by FCC rules.[16,17] In order to use the spectrum more efficiently, two polarizations are used, as explained in the introduction to satellite communications. In North America, linear polarizations are used. In many other places,

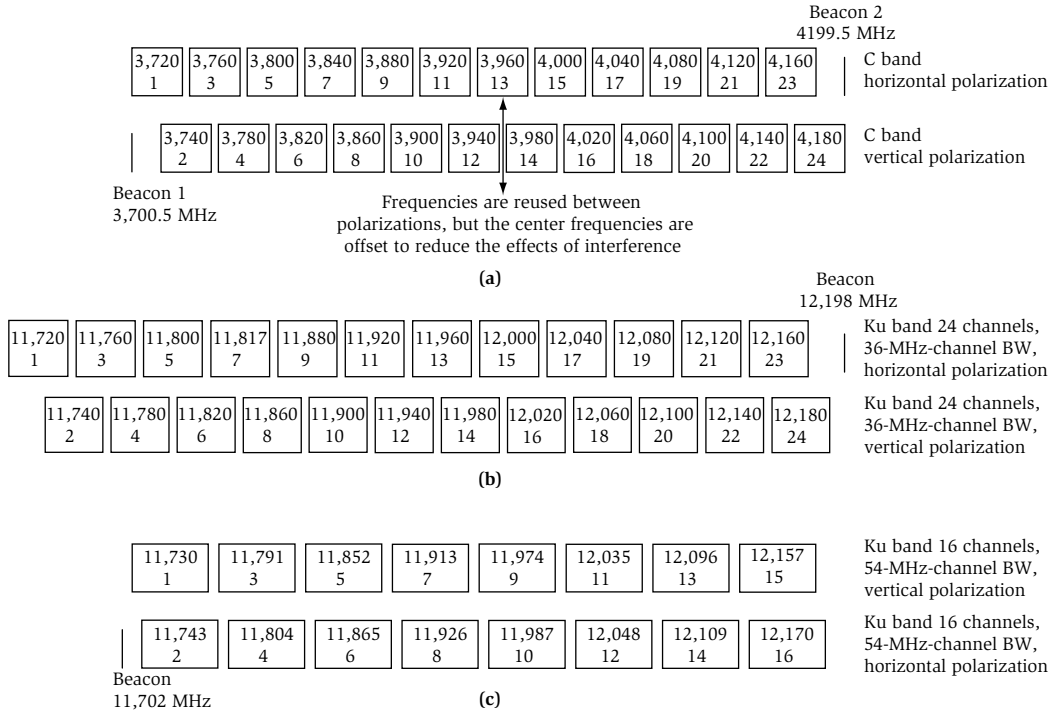


Figure 7.16 Satellite downlink frequency plans. (a) C band. (b) Plan for Ku band of AMC-1 dual-band satellite. (c) Ku-band plan for AMC-5 Ku band only.

right- and left-hand circular polarization is employed. Polarization diversity allows transmission of two signals in the same frequency band without interference. However, as a practical matter, the discrimination that we can achieve is not perfect. In order to improve the performance of the system, signals using the two polarizations are also offset in frequency from each other.

A channel is 40 MHz wide, with the occupied bandwidth of the signal approximately 36 MHz, leaving a 4-MHz guard band between channels on the same polarization. The opposite polarization is offset in frequency such that its picture carriers are in the guard band, as shown in Figure 7.16(a).

Ku-band plans are not as standardized as is the C-band plan. Figure 7.16(b) illustrates the Ku band downlink plan of the AMC-1 satellite. This is a combined C/Ku-band satellite with Ku-band channels 36 MHz wide having a 4-MHz guard band, just as in the C-band plan. Figure 7.16(c) shows another Ku-band plan, using 56-MHz-wide channels having a 7-MHz guard band. This plan, used in the AMC-5 satellite, allows one transponder to be used for two analog TV programs or a higher-data-rate digital signal.[18] Note in all plans that on top of the channel we show the center frequency, in megahertz, for that transponder, and we show the channel number under it. The satellite bands are 500 MHz wide in North America, whereas in Europe they are 800 MHz wide.

Each channel corresponds to one transponder on the satellite. A transponder works much like a cable TV signal processor, explained in Chapter 8. It is simply a frequency translator and power amplifier. Signals are uplinked to the satellite at a frequency roughly 2 GHz higher than the downlink band. On the satellite, the transponder receives and amplifies the signal, converts it to the output frequency, amplifies it, and feeds it to the transmit antenna. The *beacons* are narrow-bandwidth signals that transmit telemetry information related to the condition of the satellite. These beacons, sometimes called *telemetry* (T/M) *beacons*, are used by satellite controllers.

It is common for the downlink frequency band to be converted to a lower-frequency band at the antenna, before the signal enters the headend. This is done in the preamplifier located at the feed point of the antenna. The device that provides preamplification and down-conversion is called a low-noise block converter, or LNB (see Figure 8.16). In North America the common band to which signals are down-converted is 950–1450 MHz. In Europe, several bands are used at frequencies up to 2300 MHz. Since the signal is down-converted at the LNB, the receiver is the same for either C or Ku band.

Polarization Rotation

Even if the signals leave the satellite polarized exactly vertically and horizontally, they don't necessarily arrive at the earth so polarized. Faraday rotation will cause both signals to rotate. Faraday rotation in the northern hemisphere is clockwise looking from the satellite toward the earth station for a downlink signal. Rain will also cause rotation of the polarization. The

net effect is that the earth station antenna's ability to reject the undesired polarity is limited.

Faraday rotation is generated in the ionosphere. The ionosphere is a neutral gas containing charged particles called a plasma. In the presence of a magnetic field (from the earth), a plasma becomes an anisotropic medium, which will cause rotation of an electromagnetic signal propagating through it. Michael Faraday observed the same thing in light traveling through gyrotropic crystal with an applied magnetic field, so the phenomenon came to be named for him.

The amount of Faraday rotation depends on the electron density in the ionosphere, the magnetic field intensity, and the component of the wave traveling along the magnetic field vector. It will thus vary, as the concentration of electrons in the ionosphere varies with the season and with the time of day. The amount of Faraday rotation is inversely proportional to the square of the frequency, so it is more severe at C band than at Ku band. A typical peak value at 4 GHz is 4°. The loss of polarization discrimination at the earth station is given by the sine of the angle of rotation, so 4° results in polarization discrimination of about 23 dB. Of course, the feed may be adjusted for an average Faraday rotation, and if this is done correctly, the loss of polarization discrimination may be much less. Depolarization due to rain may be more severe in a number of locations.

Rain can cause depolarization because the raindrops are elliptical and cause amplitude and phase distortions in the two orthogonal linear polarization components of any incident polarization. Rain depolarization tends to be more severe at higher frequencies. Depending on the rain rate and the elevation angle, rain depolarization can cause the polarization discrimination to drop to between 30 and 10 dB (very high rain rates).[19]

Thus, though polarization can normally yield some 20 dB of discrimination against the undesired polarity, there are times that we wouldn't be able to expect much more than 10 dB of discrimination, thus the need for frequency offset as well as polarization discrimination.

Numerical Observations Regarding Satellite Receiving Stations

The FCC uses a 2° separation standard for satellites serving the United States. Many satellites are so spaced, mostly near odd numbers of degrees. The physical separation is referenced to the center of the earth. Since real earth stations are located at the surface of the earth, about 15% of the distance from the earth's center to the synchronous orbit, the apparent spacing is going to be larger than is the specified orbital spacing.

Cook has presented an equation showing the apparent separation between two satellites from any place on the surface of the earth:[20]

$$\cos\alpha = \frac{2C_2 - C_3\cos X(\cos\Delta L_A + \cos\Delta L_B) - 2C_1(1-\cos\beta)}{2\sqrt{C_2 - C_3\cos X\cos\Delta L_A}\sqrt{C_2 - C_3\cos X\cos\Delta L_B}} \tag{7.15}$$

where

$C_1 = 43.70532$
$C_2 = 44.70532$
$C_3 = 13.222$
$X =$ station latitude
$\Delta L_A =$ difference in longitude between first satellite and station (degrees)
$\Delta L_B =$ difference in longitude between second satellite and station (degrees)
$\beta =$ satellite separation (degrees)

From this and other equations shown earlier, we can illustrate the geometrical issues for two satellites separated by 2° for several locations in the United States. Choose the eastmost positions in the orbital arc (84° and 86°) for which the satellites would be visible at all locations illustrated. (For illustration, we're using even-numbered degrees though most satellites over the Americas are at odd-numbered degrees.) The orbital slots shown are not occupied by actual satellites, but are intended to give a feel for the range of values we might encounter at moderate latitudes. Table 7.2 shows the azimuth and elevation for earth stations located in several places, for the satellite at 84°. The table also shows the apparent separation between that satellite and one at 86° and the slant range and space loss for the satellite at 84°.

The satellite locations are toward the east end of the orbit. They are roughly south of the Florida panhandle, one south of the eastern part and one south of the western part of the panhandle. The slant range from the farthest earth station (Alaska) is about 14% greater than the closest (Florida). However, the space loss, caused by spreading of the beam from the satellite, varies only by about

Table 7.2 Earth Station Geometry and Space Loss for Representative Locations in the United States for Desired Satellite at 84° and Interfering Satellite at 86°

	Lat. (deg)	*Long. (deg)*	*Az. (deg)*	*El. (deg)*	*App. Sep. (deg)*	*Slant range (statute miles)*	*Space Loss (dB)*
Equation:			(6.8)	(6.9)	(6.15)	(6.10)	(6.11)
Location							
Alaska	67	150	112	0.4	2.02	25,867	197
Hawaii	20	155	96.7	9	2.06	25,270	197
Washington State	47	120	135	26	2.16	24,239	196
Illinois	45	90	172	38	2.22	23,581	196
Florida	27	82	184	58	2.30	22,742	196

1 dB. A more dominant factor in determining station performance is the pattern of the antenna on the spacecraft. Typically, the EIRP in a particular direction will vary by a few decibels compared with that in another direction.

The apparent separation between the satellites is always at least as great as is the actual separation (as seen from the center of the earth). This may not be intuitive: you might expect that an earth station far from the subsatellite point (the point on the surface of the earth directly under the satellite), and near the equatorial plane, might see less separation than 2°. In this example, Hawaii is the station you might expect to experience the small difference in apparent separation. However, because the earth is round, Hawaii is "below" the normal plane at the subsatellite point, leading to more apparent separation.

7.3.5 Earth Station Design Example

A brief example of the parameters of an earth station design can be obtained from the preceding numerical examples. Assume a satellite at a longitude of 84°, as in Table 7.2 (no real satellite is assigned to this location). Compute the azimuth and elevation, which tell how to point the antenna. The space loss is computed once the slant range is determined. In order to compute the performance of the earth station, it remains to determine the carrier-to-noise ratio at the receiver. The first step is to obtain the EIRP of the satellite at the chosen location. This is obtained from the satellite operator. A typical C band number might be 34 dBW. Assume the Washington State location, with a path loss of 196 dB, and an earth station with an effective G/T of 20 dB/°K. Finally, assume that the receiver manufacturer specifies an IF noise bandwidth of 32.5 MHz. Equation (7.14) now becomes

$$\begin{aligned}\frac{C}{N} &= EIRP - L_s + \frac{G}{T} + 228.6 - 10\log(BW_n) \\ &= 34 - 196 + 20 + 228.6 - 10\log(32 \times 10^6) \\ &= 11.55\,\text{dB}\end{aligned}$$

The receiver will receive a signal having a carrier-to-noise ratio of 11.55 dB. If the S/N versus C/N curve is as shown in Figure 7.14, the expected video–signal-to-noise ratio will be on the order of 46 dB. The receiver is operating some 4 dB above threshold, before allowance for fade. Rain fade, antenna misalignment, polarization errors, foliage blockage, and other factors will reduce the actual carrier-to-noise ratio. Also, nothing has been allowed for terrestrial interference, interference from adjacent satellites, or the uplink (usually much better than the downlink). A prudent operator will allow several decibels for such.

7.4 Other Methods of Receiving Programming at Headends

By far the most common methods of receiving signals at headends is either by off-air antennas or by satellite. However, other methods are in use, and may even increase in importance in the future. These are covered briefly.

7.4.1 FM Terrestrial Microwave

In earlier days of the industry, it was very common to use terrestrial FM microwave to bring signals to the headend from distant television stations. Many of these links have now been replaced with fiber-optic cable or with closer television stations, but some links remain in operation. The equipment was the same as that used at the time by the telephone industry for microwave transmission, with the exciter changed to handle video. The standards used for modulation were very similar to those used for satellite work, so the concepts discussed earlier are valid for terrestrial microwave also. Propagation issues relating to microwave transmission are similar to those discussed in Chapter 14.

7.4.2 Direct Feed from Television Stations

Because of the perils of the radio path, it is sometimes not possible to deliver high-quality signals to a headend using pickup off the air. A moderate number of television stations now are using direct feeds from their studios to the cable headend. Typically, the baseband output of the master switcher, which supplies the broadcast transmitter, is taken as the input to the cable system.

A related activity today is cooperation between a local television station and one or more cable systems to produce local programs. Typically, the news department of a television station prepares a special cable edition of their newscast, or the off-air newscast is repeated on the cable. Other programming may also be involved, produced by either the television station or the cable operator.

Several methods are used to transport signals to the headend. Today, the most popular methods are analog using either coaxial cable or fiber optics. The trend is toward digital transmission using fiber optics. Some of the more common options are listed next.

Baseband Transmission

It is possible to transmit baseband signals several miles on coax, but the signal must be carefully conditioned to compensate for the considerable rolloff in response of a coax cable in the baseband spectrum. Also, hum pickup is a possibility that has to be dealt with. For stereophonic sound, the audio will have to be picked up off the air, brought down to the 4.5-MHz intercarrier frequency (or perhaps the 41.25-MHz sound IF), and combined with the video after modulation. Alternatively, the

sound signal could be modulated onto a 4.5-MHz subcarrier at the TV station, and the sound subcarrier combined with the video for transmission. Most higher-end modulators can handle signals delivered in this fashion.

AM Supertrunk

It is possible to locate a cable television modulator at the TV studio and transport signals on a dedicated coax or fiber-optic path to the headend. At the headend, the signal is supplied to a processor of the same type used for off-air signals, which converts its frequency to the final cable channel and connects the signal to the outgoing cable plant. This is probably the oldest technique in use. If stereo transmission is to be employed, then a stereo encoder must be located at the studio so that BTSC stereo is transmitted from the studio. At the headend, most cable processors will handle BTSC stereo signals without modification. However, some operators have picked up the off-air signal using a second processor, and combined the sound IF from that processor with the picture IF from the processor handling the video coming from the studio. Engineers must be careful in doing this so that they don't exceed the frequency tolerance of the intercarrier frequency (see Chapter 8). The FCC requires that the intercarrier frequency error not exceed ± 5 kHz on the cable (the broadcast tolerance, respected by high-end modulators, is ± 1 kHz). If one processor is used to take an incoming picture carrier, convert it to a second channel, and combine with a sound carrier coming from an independent source, it is not too likely that the frequency tolerance will be met. Engineers must do custom-tuning of the oscillators involved in order to make this happen.

AM Sublow Band

Often if a cable operator wants to get programming originating in the community back to the headend, he or she will *backhaul* it on the sublow portion of a trunk cable. Many cable systems carry local sports events, government meetings, and government-sponsored information. The backhaul is frequently done by installing two-way amplifiers in the required portion of the plant. A sublow modulator is used at the point of origination, and at the headend the signal is converted to the downstream channel, again using a processor. This is essentially the same as the technique just discussed, except for the use of the sublow spectrum on the necessary portion of the plant.

By common convention, the industry has divided the return band into 6-MHz "channels," which are designated T channels. Table 7.3 shows the traditional designations used. Generally, the higher-frequency channels provide the higher-quality signal, due to noise on the return path. As shown in Chapter 16, frequencies generally below 15 or 20 MHz tend to experience a lot of interference on the return path. Thus, higher frequencies are preferred, especially for analog video, which tends to be more susceptible to interference than is digital transmission.

Table 7.3 Sublow Channel Designations

Channel	*Picture carrier (MHz)*
T-7	7
T-8	13
T-9	19
T-10	25
T-11	31
T-12	37
T-13	43

One factor, though, forces us to not put the return video too high in the spectrum. As shown elsewhere, the diplex filters used in two-way distribution amplifiers produce group delay near the crossover frequency. If the signal must be transported through too many amplifiers, the group delay may add to an unacceptable number. The usual countermeasure for this is to reduce the frequency of the analog transmission.

FM Supertrunk

A similar technique is to use FM transmission on a dedicated cable. This was popular a few years ago since the quality signal delivered to the headend could be quite good. The same conditions apply to audio as described before.

Digital Transmission with No Data Compression

Equipment is available today to allow noncompressed digital transmission to take place over a fiber-optic cable. The systems are low in cost but require a relatively high data rate for each channel. Data rates of 120–240 Mb/s per video channel are not uncommon in these applications. Several video signals can be multiplexed, with the total data rate being 2.4 Gb/s or higher. The technique is also used to interconnect headends, where it is desired to transport several high-quality signals between headends. A dedicated fiber is usually used, though wave division multiplexing (WDM) is beginning to be practiced today.

Digital Transmission with Moderate Data Compression

Equipment is available to allow moderate compression in a transmission path. Generally, this technique will be more costly than use of noncompressed transmission, but it may be preferred if there is a scarcity of bandwidth.

Compression is based on vendor proprietary compression schemes, most of which use *linear predictive coding* (LPC). In an LPC system, an algorithm is used to predict the content of the next picture (usually either a field or frame) based on past data. The encoder includes a decoder, which works identically to the one at the receiving end. The predicted output of the decoder is compared with the picture before encoding. The difference is an error signal, which is transmitted to the receiver. Because TV pictures tend to have a lot of commonality between frames, the amount of information transmitted is much less than if no compression were used. Data rates of around 45 Mb/s can yield excellent-quality pictures and are routinely used for backhaul of programs to broadcasters. Data rates in the neighborhood of one-half of this have met with limited acceptance.

Digital Transmission with Extreme Data Compression

This usually means some form of MPEG or MPEG-derived compression algorithm based on the *discrete cosine transform* (DCT) and possibly other aspects of MPEG compression. Because of the cost, no known applications exist today that utilize full MPEG compression for point-to-point transmission of video though this may happen in the future. However, it is not that expensive to implement just the DCT and maybe some other modest data reduction techniques. Data rates of under 20 Mb/s for broadcast quality have been reported. Chapter 3 includes more information on compression.

7.4.3 Local Origination

Finally, local origination (LO) is practiced in many systems. LO is done from studios operated by the cable company or sometimes by other parties. Typically, the studios will be similar to those at broadcast stations, except that the amount and quality of equipment is less.

Much programming originates on tape. Common tape formats in use include consumer VHS (very low end), super VHS, and 3/4 inch, in increasing order of quality and cost. Better operations employ time base correctors to compensate for the unavoidable timing errors that develop because of inconsistent tape transport speed. Computers may be used to control the tape machines, or they may be controlled manually.

The leading edge in local origination storage is the use of video file servers. Programs are digitized and compressed. The compressed program is stored on a large file server, not unlike those used in business applications to store massive amounts of data. At the programmed time, the file server "plays" the file to an MPEG-to-NTSC converter (for analog channels), which converts the signal back to analog. This type of system saw its earliest successes in the commercial insertion (ad insertion) market, is now used for playing out long-form programs.

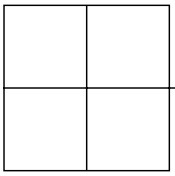
7.5 Summary

This chapter has discussed the underlying technology used to bring signals into the headend. The vast majority of signals are received today from satellites, with the next largest number being picked up off the air. Other techniques are used and are outlined briefly here. In the next two chapters, we will discuss headend-related concepts beyond getting signals into the headend. Chapter 8 will cover handling signals in individual pieces of equipment in the headend, and Chapter 9 will cover handling the signals outside those pieces of equipment.

Endnotes

1. Richard C. Johnson et al., *Antenna Engineering Handbook*. New York: McGraw-Hill, 1984, pp. 4–11.
2. Mark J. Wilson, ed., *The ARRL Handbook for the Radio Amateur*. Newington, CT: American Radio Relay League, 1988, p. 17.2. Regrettably, this figure seems to have been removed from recent editions.
3. Computed from Equations (12) and (14) of International Telephone and Radio Corporation, New York, *Reference Data for Radio Engineers*, 4th ed., 1967, pp. 676–677.
4. Scientific-Atlanta, Inc., Atlanta, GA, *CATV Data Book*, available at www.sciatl.com.
5. Richard C. Johnson et al., op. cit., pp. 29–17ff.
6. Thomas Milligan, *Modern Antenna Design*. New York: McGraw-Hill, 1985, p. 383ff.
7. Scientific-Atlanta, Inc., Atlanta, GA, *CATV Data Book*, rev. 6, August 1990, p. 77.
8. Marty Barringer, Radio Wave Propagation, *Engineering Handbook*, 7th ed. Washington, DC: National Association of Broadcasters.
9. Ibid.
10. Rick Barker and John Friesz, Earth Station Antennas, RF Considerations. Scientific-Atlanta Communications Symposium, 1983.
11. Ibid.
12. F. M. Fonda, *Earth Station Geometry*. Scientific-Atlanta Communications Symposium, 1983.
13. Jack Sanders, Satellites: The Birth of an Industry. *Installer/Technician*, May 1988, p. 15.
14. J. Searcy Hollis, *Satellite Link Analysis*. Scientific-Atlanta Communications Symposium, 1983, p. 12.

15. European Telecommunications Standards Institute, *Digital Broadcasting Systems for Television, Sound and Data Services: Framing Structure, Channel Coding and Modulation for 11/12-GHz satellite services*, ETS. 300 421, December 1994. Available at http://www.etsi.org/.
16. 47 CFR sec. 26.211. U.S. Government publication, available at Government Printing Office bookstores in major cities.
17. J. Searcy Hollis, Principles of Satellite Communications. Scientific-Atlanta Communications Symposium, 1983, p. 25.
18. www.ses-americom.com
19. Rick Barker, *Re-Use Earth Station Antennas*. Scientific-Atlanta Communications Symposium, 1983.
20. James H. Cook Jr., *Interference Analysis*. Scientific-Atlanta Communications Symposium, 1983. (The equation is in Appendix A.)



Chapter 8
Headend Signal Processing

8.1 Introduction

This chapter deals with the electronic devices most commonly used in headends to format signals for transmission on a cable plant. These devices include signal processors, RF devices used to take television signals off the air or from an incoming cable and put them on the cable. Also included are modulators, miniature transmitters that accept baseband audio and video signals and convert them to a channel. Demodulators do the opposite, converting an RF signal back to baseband video and audio. Stereophonic encoders convert left- and right-channel audio into the composite stereo format used in television transmission. Earth station receivers convert signals coming from communications satellites to baseband audio and video. Often these are used with descramblers that deal with the proprietary scrambling format currently used with most analog satellite transmissions.

Block diagrams and descriptions of the operation of the equipment are presented, along with information regarding the proper application of the hardware. The block diagrams are representative of commonly available equipment. They are intended to illustrate how the functions are accomplished; they are not intended to represent particular pieces of equipment. Variations will be encountered in the field. Some equipment may have controls that don't work as do the controls shown here, as a consequence of differences in block diagrams.

Chapter 7 covered inputs to the headend from various RF sources. Chapter 9 will cover methods used to couple signals from the equipment covered in this chapter to the cable. You may wish to review the fundamental composition of baseband and RF signals in Chapters 2, 3 and 4.

8.2 Signal Processors

RF signal processors are normally used to transfer incoming VSB-AM signals from off-air antennas or incoming cables to the cable plant. The processor can change the channel (frequency) of the signal and the relative sound carrier level. The signal is bandpass filtered to remove any undesired adjacent channel power.

Automatic gain control is employed to stabilize the output level. Optionally, a standby carrier can be substituted when the incoming signal goes away. Switching of IF sources can be done to maximize use of a channel.

8.2.1 Input Section

Figure 8.1 illustrates a typical block diagram of a processor. A number of variations are possible. Many modern processors are frequency agile. Differences between agile and nonagile processors are pointed out as appropriate. The incoming RF signal is supplied to a bandpass filter, BPF1, which is tuned to the incoming channel. This bandpass filter is typically of fairly low order and has only a little adjacent channel rejection. Its primary purpose is to reduce off-channel power to minimize distortion in the input circuitry. It, along with BPF2, provides image rejection, as described in Chapter 9.

The signal passes through a variable attenuator, AT1, which is part of the delayed AGC circuit. "Delayed AGC" is a somewhat unfortunate term because it has nothing to do with a time delay. Rather, the term *delayed* refers to the fact

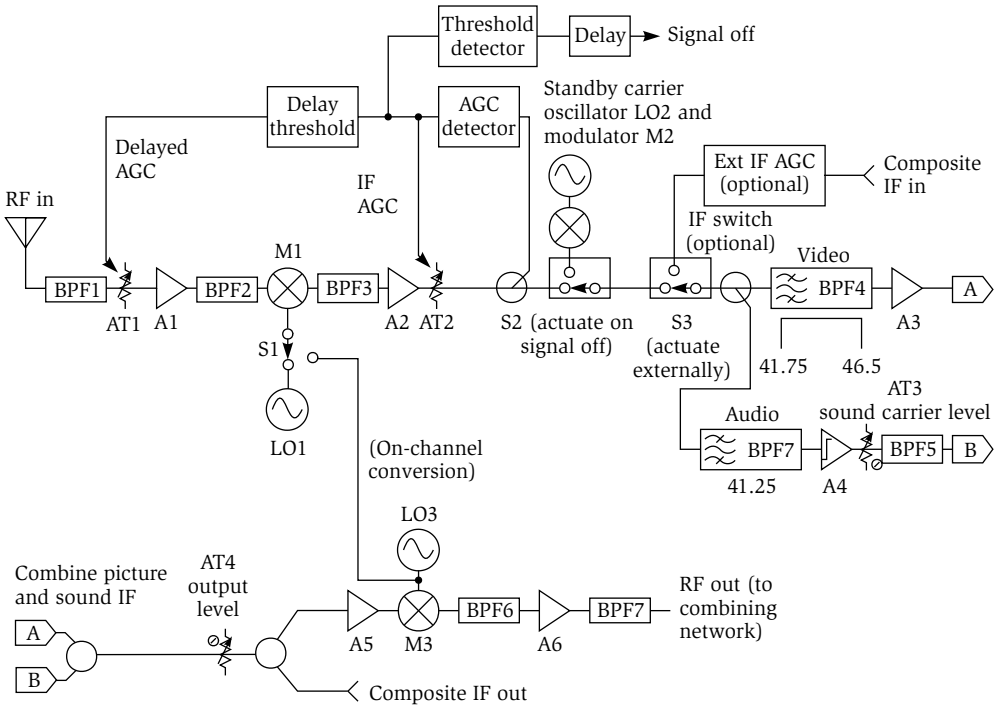


Figure 8.1 Signal processor block diagram.

that AT1 is at minimum attenuation (the front end is at maximum gain) at low signal levels. As the signal level increases (above a predetermined threshold), it becomes necessary to reduce its amplitude before it reaches RF amplifier A1 and mixer M1. Failure to reduce the amplitude would result in unacceptable distortion. At lower signal levels, the gain of A1 is needed in order to overcome the noise introduced in the mixer and following circuitry.

8.2.2 IF Section

The signal is converted to the intermediate frequency (IF) in mixer M1. The commonly used North American IF is 45.75 MHz for the picture carrier and 41.25 MHz for the sound carrier.* Note that the spectrum is inverted at IF compared with the spectrum at RF, where the sound carrier is at a higher frequency than is the picture carrier. This is a consequence of using a high-side local oscillator (LO); that is, the frequency of the local oscillator is higher than is the RF frequency of the signal being converted. Use of high-side oscillators helps control spurious responses from the mixing process.

Following mixer M1 is another bandpass filter, BPF3, tuned to the IF band, normally 41 to 47 MHz for NTSC signals. Amplifier A2 provides gain to help overcome the insertion loss of later circuits. The IF AGC attenuator is AT2, which is adjusted such that the input to the AGC detector is constant regardless of the incoming signal level. Additional amplification may be provided, and at some point in the signal chain, a sample of the signal will be routed to the AGC detector. The output voltage of the AGC circuitry changes as the signal into the AGC detector increases. This change causes the attenuation of AT2 to increase, compensating for the increasing level of the incoming signal.

If the signal level increases sufficiently, the IF AGC voltage reaches the delay threshold, above which the gain of the input attenuator, AT1, is reduced to effect stabilization of the IF signal level. Depending on the design of the AGC system, the IF attenuator, AT2, may continue to drop in gain, or it may not. Usually, the best carrier-to-noise performance of the processor is achieved at input signal levels just below the delayed AGC threshold. If sufficient level is available, this is the preferred operating point for the processor.

Returning to the main IF signal path, as illustrated, the signal next is supplied to switch S2, which selects the standby carrier upon loss of input signal. The standby carrier was originally employed to prevent TVs from showing snow when a signal goes off the air. It is also employed if the output channel is used as a pilot for the distribution system. A pilot signal is used to set the gain of distribution systems, and if it disappears when a TV station goes off the air, the operating point of the distribution system will be upset.

---

* PAL countries use 38.9 MHz for the picture carrier IF and 33.4 (PAL-B/G) or 32.9 (PAL-I) MHz for the audio IF. In China, 38 MHz is used for the picture carrier IF.

Some processors include the ability to modulate the standby carrier with pseudosync pulses, a low-duty cycle 15.734-kHz square wave that simulates normal TV modulation. This is necessary to support distribution AGC systems that need sync tips to operate, rather than a CW signal.

Switch S2 is normally activated by a drop in the IF AGC voltage sufficient to indicate that the input signal level is lower than a minimum needed for useful performance. Following switch S2 is switch S3. This switch is optional and is used when it is desired to switch to an alternative IF signal. This is done, for example, when the primary signal leaves the air and it is desired to substitute another signal on the same channel. For example, the cable operator may carry a local television station that goes off the air at night. It is desired to substitute another signal on the channel when the local station goes off the air. The alternative signal, processed in another modulator or processor, is supplied at the "Composite IF in" connector. Optionally, a separate AGC circuit may be used to normalize the level of the alternative IF signal. In this example, switch S3 would be connected to the IF AGC signal such that when the input signal drops, S3 is activated, selecting the alternative IF signal.

The picture and sound carriers are separated in the example processor. Aural IF, at a frequency of 41.25 MHz, is routed to a bandpass filter, BPF7, which eliminates any power associated with the picture signal or other channels. Following BPF7 is a limiter, A4, which functions identically to the limiter in an FM radio. Its output is constant regardless of the input signal level, within a reasonable level range. Following A4 is a manually adjusted attenuator used to adjust the sound carrier level. The sound carrier is normally adjusted to be −15 dB from the picture carrier level. A bandpass filter following AT3 removes harmonics of the sound carrier, which are introduced in the limiting process. The sound IF is then combined with the picture IF.

An alternative sound IF circuit uses a notch of variable depth to adjust the sound carrier level. This circuit works but cannot remove any residual AM on the sound carrier. Further, if not adjusted properly, it can add FM/AM noise, which hurts operation of many scrambling systems.

The visual signal at IF is supplied to BPF4, which provides the bulk of the selectivity of the processor. We describe the response of this filter later. After amplification, the visual IF is combined with the sound carrier.

8.2.3 Output Section

After combining, the composite IF signal is applied to output level control AT4. The location of this control in the signal chain allows it to simultaneously adjust the picture and sound carrier levels. After level control, the composite IF signal is split. One side of the splitter is supplied as the "Composite IF out" of the processor, to be used if needed as the alternative IF input to another processor or modulator. The other output of the splitter is amplified in A5 before being supplied to mixer M3, the output mixer. It is mixed with the output of local

oscillator LO3. The signal is now on the assigned output channel. Bandpass filters BPF6 and BPF7 eliminate the image of the second conversion process and any spurious signals coming from mixer M3.

Amplifier A6 increases the signal level to the desired output level. Its performance is critical: any third-order distortion it introduces will generate frequencies too close to the output channel to be rejected by BPF7. The most common distortion produced is a third-order beat between the picture and sound carriers. This beat, appearing 4.5 MHz below the picture carrier, is derived from twice the picture carrier less the sound carrier. Another beat appears 4.5 MHz above the sound carrier but is lower in amplitude.

If the output is *frequency agile*, meaning it can be set to any of a number of channels, the output section shown is replaced with a dual conversion output section, with a first IF above the maximum frequency the processor can output. The IF is filtered to remove the image, and then it is down-converted to the output channel. Sometimes up-conversion is done externally to the modulator; this technique, along with more information regarding up-conversion, is covered in Section 8.4.1. The mechanism that creates images is covered in Chapter 9.

8.2.4 Local Oscillators

Normally, LO1 and LO3 are fixed, crystal-controlled oscillators in a fixed channel processor. If the processor is agile on the output channel, LO3 and M3 will be replaced with a dual conversion output converter, which converts the IF signal to a high frequency (typically in the high hundreds of MHz), where it is filtered and down-converted to the output channel using yet another mixer and a phase locked local oscillator.

A processor with an agile input will have the input section replaced with an agile oscillator and tracking filter. In higher-quality processors, the front end will be double conversion, with the second IF at the normal TV band of 41–47 MHz.

If the processor is used as an on-channel processor, with the input and output on the same channel, then it is desirable that the output signal be phase coherent with the input signal to avoid beat products. To effect this phase coherency, local oscillator LO1 is not used for on-channel conversion. Rather, a sample of the output local oscillator, LO3, is supplied to mixer M1. Thus, the same local oscillator is used for both down- and upconversion. This may be shown to ensure phase coherency. Switch S1 is thrown to the alternative position for on-channel conversion.

If it is desired that the output be phase locked to an external reference signal at the frequency of the output channel, then LO3 will be replaced with a phase locked circuit.

8.2.5 Processor Frequency Response

Figure 8.2 illustrates the visual channel IF response of the processor, overlaid with the spectrum of signals that may be present. An important function of a

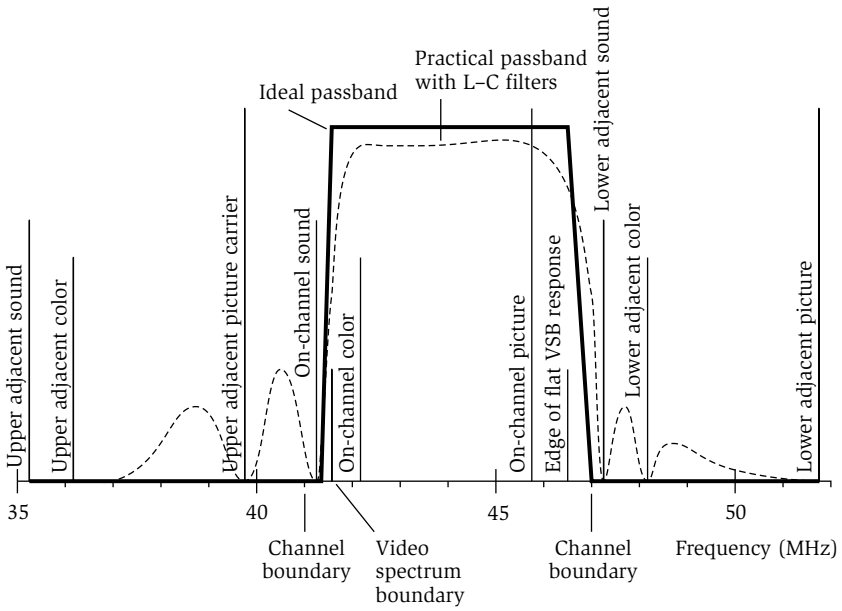


Figure 8.2 Processor visual channel IF response.

processor is to provide off-channel rejection. This is important in cases where the processor is used to process signals coming in from another cable. For example, it is sometimes practiced to route signals by AM microwave or cable to a hub, where they are reprocessed and supplied to subscribers. In this case, adjacent channel power coming into the processor must be rejected.

Also, if a station is being received off the air along with an adjacent channel from another city, the adjacent channel must be rejected before the signal is supplied to the cable. It is accepted that, in order to be invisible, an undesired signal should be 60 dB below the level of a desired signal. For specification purposes, this is the desired adjacent channel rejection of a processor. In the case of incoming cable signals, adjacent signals will be at about the same level, so 60-dB adjacent channel rejection is adequate. If an adjacent channel has a greater signal level than the tuned channel, as could be the case in off-air processing, then more rejection may be required. This is considered to be handled best with external selectivity rather than to burden all processors with more rejection capability.

Figure 8.2 illustrates the ideal processor passband as a heavy line and a typical passband as a dashed line. The location of the relevant carriers is shown at IF, where the picture carrier is at a higher frequency than the sound carrier. Chapter 8 shows why the spectrum is inverted from that broadcast. Ideally, the

visual IF passband would pass signals from the edge of the visual spectrum, 4.2 MHz from the picture carrier, to the edge of the flat portion of the vestigial sideband region, 0.75 MHz above the picture carrier at IF.

This ideal spectral response may be approximated if a surface acoustic wave (SAW) filter is used as the IF filter (BPF4 in Figure 8.1). This filter sets the overall visual IF response of the processor for all practical purposes. Some processors employ L–C (inductor–capacitor) filters for this application. Practical L–C filters are unable to achieve the ideal passband shape. Bandpass sections are combined with traps, which provide high rejection over a very narrow band of frequencies. These traps are placed on the adjacent carrier frequencies as shown. Most of the power in an NTSC signal is concentrated at the picture carrier, the sound carrier, and the color carrier. The power density at other frequencies in a channel is relatively low. Because of this characteristic, it is possible to specify 60-dB rejection even if that much rejection is not available at all frequencies. The specification of 60-dB rejection is interpreted as meaning that power in an equal-level TV signal appearing in an adjacent channel is more than 60 dB below the on-channel picture carrier.

Even when a SAW filter is employed, it may be necessary to supplement its response with some traps. Practical SAW filters exhibit ultimate rejection of 40 to 50 dB, against a requirement of up to 60 dB depending on the frequency. Traps may be needed to achieve this last little attenuation.

L–C filters exhibit group delay. This means that signal components passing through the filter near either edge are delayed with respect to signal components passing through near the middle of the passband. The delay is a natural consequence of the attenuation of the stopband and is predicted from the mathematics describing the response of the filter. If the transition from passband to stopband occurs in a smaller frequency band, the group delay is greater.

Group delay causes components of the visual signal that occupy frequencies near the edge of the passband to propagate through the filter more slowly than do components near the center of the passband. This causes undesirable distortion of the video signal. To remedy this, another filter, called a *delay equalizer*, is incorporated with the bandpass filter. The delay equalizer must be designed for the bandpass filter with which it is to be used. It adds more delay to signals at the center of the passband, complementing the delay response of the bandpass filter.

A delay equalizer is necessary when an L–C filter is used as the bandpass filter, but not when a SAW element is used. Because of the way a bandpass filter is realized using a SAW device, group delay is inherently low.

8.2.6 Digital Off-Air Processing

To meet some early needs to carry off-air digital signals on cable plant, a simple system was used. An analog signal processor was retuned to widen the IF passband shown in Figure 8.2 such that it would pass the 8-VSB signal picked up off the air. The AGC was bypassed by placing the processor in the manual mode. Signals were up-converted to an off-air UHF frequency (which differs

from the cable UHF frequency plan by 2 MHz, as explained in Chapter 9). This is not particularly efficient method of carrying off-air signals, but it sufficed in the early days, when adequate spectrum was available on the cable plant.

As shown in Chapter 4, off-air digital signals use 8-VSB modulation with a transport stream payload data rate of 19.393 Mb/s. The data rate of a 64-QAM video signal on cable (after removing transmission overhead) is 26.97 Mb/s; for 256-QAM it is 38.811 Mb/s. You can see that a 256-QAM channel can accept two complete 8-VSB modulated channels. Furthermore, in many cases a cable operator may want to carry some but not all of the data in the off-air channel, freeing up more spectrum. When cable operators carry off-air digital channels it is important to minimize the spectrum required, so there is a strong need to combine two off-air digital channels into one cable channel. This can be done with a *digital television processor*.

Figure 8.3 illustrates digital television processing, the digital equivalent to analog signal processing. A digital signal is recovered off-air in tuner #1 and supplied to the PSIP processor and remultiplexer (remux). As explained in Chapter 3, digital video signals include the *program and system Information Protocol* (PSIP), a series of tables that, among other functions, tell the receiver what programs are included in the multiplex (i.e., the 6-MHz channel carrying several programs) and how to find them. These tables must be modified if some programs are removed from the multiplex and others are added from another off-air channel or other source.

In the configuration shown, a second program stream may be added from either another off-air channel via tuner #2 or from another signal source. This program transport stream is supplied to the PSIP processor and remux. Its PSIP tables are merged with those from the first program stream and multiplexed into a higher-bandwidth stream that is supplied to a 256-QAM modulator. The modulator may include up-conversion to the frequency the combined transport stream occupies on the cable plant, or the up-conversion may be separate. Since each program stream included in each of the two input streams shown may be transmitted in

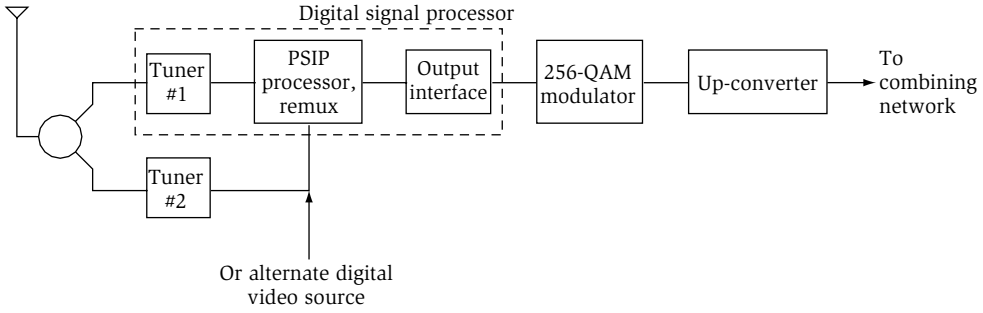


Figure 8.3 Digital television processing.

variable-bit-rate format (VBR — explained later), there are some issues that must be addressed relating to the amount of data that is carried in the combined transport stream. In the case of combining two off-air channels, there is enough bandwidth to carry two complete channels. In many cases, however, only the HD elements are desired, and the remultiplexing process is capable of dropping not only the other television services but other data services as well. This would leave two off-air HD signals, with perhaps enough remaining bandwidth for another SD service of the operator's choosing. The remultiplexing process must also derive and recombine all of the PSIP information from the desired incoming sources.[1]

8.3 Modulation

Figure 8.4 illustrates the video baseband, IF processing, and RF processing found in a typical modulator. (Upcoming Figure 8.5 shows the audio circuitry.)

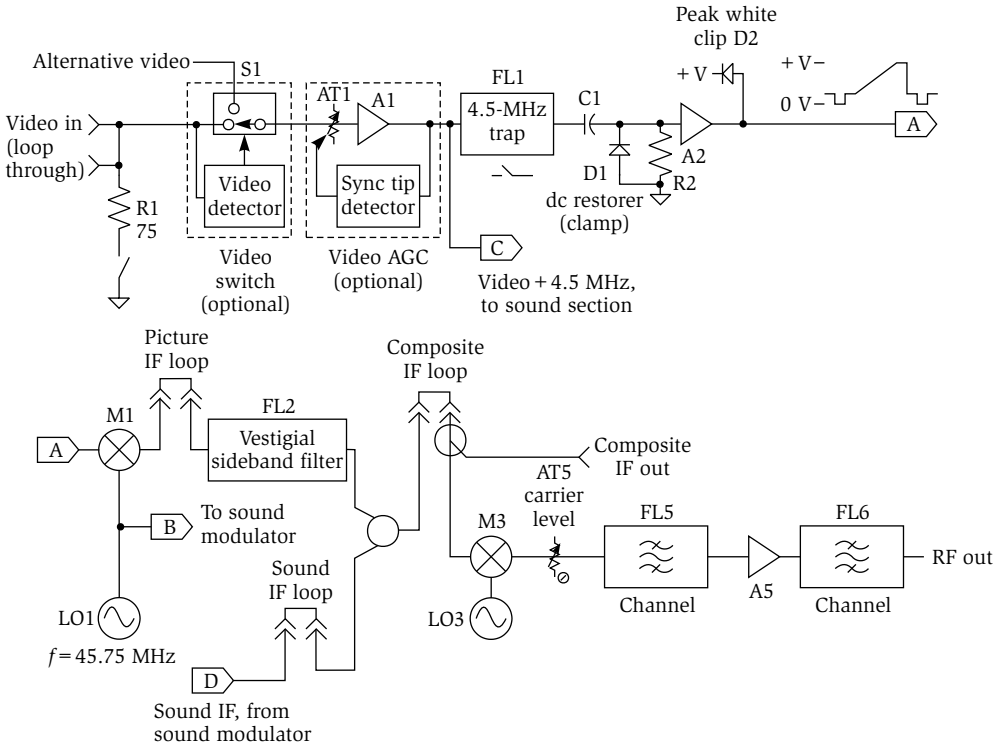


Figure 8.4 Modulator block diagram, baseband and RF.

Modulators are used with many more channels than are processors since the majority of channels are programmed from satellite-distributed programming or local sources. The output circuitry of the modulator is identical to that of a processor, but the front end is significantly different.

8.3.1 Video Switch and AGC

Video is introduced into the modulator in the top left-hand corner of Figure 8.4. An optional loop-through facility is shown on the input. Where a video source must be supplied to several places, it is possible to loop the video through from one piece of equipment to another by using a high input impedance in each. A termination is used only at the last piece of equipment in the signal chain. Section 8.9.1 covers this topic in more detail.

As shown, a video switch is optionally available to allow selection of an alternative video source. The switch may be activated manually, from an automation system, or automatically upon loss of the primary video.

Another option available with some modulators is a video AGC circuit. Typically, the amplitude of the sync tip is measured. If it is in error, attenuator AT1 is adjusted until the sync amplitude is normalized. It is assumed that the ratio of active video signal to the sync tip is correct. The proper level of video depends on the video content, which cannot be known at the modulator. However, the sync tip amplitude is well known. An error in sync tip amplitude is indicative of a gain error in video-processing circuitry. By correcting the sync tip amplitude, the active video amplitude is corrected.

Although the ideal correction strategy for the video AGC is to correct sync tip amplitude, a few modern modulators feature additional correction strategies. It is a fact of life that certain video signals handled in headends do not conform to the expected characteristics of amplitude and sync-to-video ratio. In order to deal with these signals, some video AGC circuits include a mode in which the amplitude of the sync tip to peak video signal is normalized rather than the sync amplitude. This mode should not be used except under certain conditions. If the video level drops because the picture intentionally goes dark, this mode would cause the sync level to increase and the video to be stretched to white unnaturally. However, the mode might be useful for video sources that are always expected to include peak white, but that are not always expected to be at the correct amplitude.

8.3.2 Sound Subcarrier Trap

Following the video AGC is a trap tuned to 4.5 MHz. In some cases, video is supplied combined with sound on a 4.5-MHz subcarrier. This is done, for example, when video is supplied via a terrestrial FM microwave link. The base-

band video and subcarrier sound must be separated before modulation. The aural signal is coupled out through flag C (to Figure 8.5). The aural subcarrier is removed in FL1, a 4.5-MHz trap. Frequently, the trap is switchable because it does introduce some group delay. It may not be needed, and should not be used, if the incoming video doesn't contain significant power at 4.5 MHz.

Under some conditions other than the existence of a 4.5-MHz subcarrier, use of the trap may be desirable. If the video contains significant power at 4.5 MHz owing to distortion or other processing, that power would appear as noise to the sound subcarrier. This can adversely affect the sound–signal-to-noise ratio, and in certain circumstances, could affect descrambling (see Chapter 21).

8.3.3 DC Restoration and Peak White Clip

From the trap, video is supplied to a dc restoration circuit. Video signals may be interfaced between pieces of equipment with dc restored or ac coupled. As video content changes, the average signal level, and hence the dc value of the signal, changes. However, since it is required to transmit modulated signals with a fixed sync tip amplitude, the dc level must be reestablished before the signal is modulated onto the RF carrier.

A number of circuits have been used for dc restoration. One of the simplest is shown in Figure 8.4. If any portion of the video signal tries to drop below ground, diode D1 conducts, charging C1 positive at the right end. Since the video signal voltage is all positive with respect to the sync tip, when the charging of C1 is complete, the sync tip will be at ground level (ignoring the diode drop in D1), with all other elements of the video signal being positive with respect to it. This dc level is maintained to the actual point of modulation. Resistor R2 is used to discharge C1 so that if the sync tip rises too high the capacitor will discharge until D1 again conducts on sync tips.

After the video signal is clamped, it is applied to a peak white clipper. Visual carrier modulation is strictly limited to 87.5% in the NTSC system. If the carrier modulation is too deep, the carrier will at some point be cut off, or very nearly so. Because of the way sound is demodulated (see Sections 8.5.5 and 9.5), carrier cutoff will cause intolerable sound buzz. In order to prevent this, many modulators include peak white clip circuits. If the video voltage exceeds voltage $+V$ in Figure 8.4, then diode D2 conducts, holding the voltage on the video line constant.

8.3.4 Modulation Process

The clamped and clipped video signal is coupled to modulator M1. It is no mistake that the modulator is shown modeled as a mixer with the multiplication sign. Amplitude modulation is a multiplicative process, as is frequency mixing. The circuits used for both amplitude modulation and mixing are essentially

identical and perform a multiplication process. Chapter 2 provides background on amplitude modulation theory.

Conventional amplitude modulation of a carrier by a sine wave is expressed by the equation

$$\begin{aligned} V_0 &= \cos \omega_c t(1 + M \cos \omega_m t) \\ &= \cos \omega_c t + \frac{M}{2} \cos(\omega_c - \omega_m)t + \frac{M}{2} \cos(\omega_c + \omega_m)t \end{aligned} \tag{8.1}$$

where

ω_c = carrier frequency

ω_m = modulating frequency

M = percent modulation of the signal (though TV signals use a different expression for modulation)

From the first form, we can see that the envelope (represented by the term inside parenthesis) never goes negative. (We can show that if the 1+ term is omitted, the carrier is suppressed, which is done when we modulate the color subcarrier.) The second form is obtained by trigonometric expansion of the first. It shows that the frequency domain includes the carrier and two other frequencies, spaced above and below the carrier by the modulation frequency. These upper and lower modulation products are called *sidebands*. The information contained within the modulated carrier is carried redundantly by the upper and lower sidebands.

A video signal is much more complex than a sine wave. The single sinusoida modulation term $\omega_m t$ is replaced by a sum of sine waves of different frequencies, which taken together, define the video waveform (see Chapter 2). The resulting modulated signal has a series of sidebands above and below the carrier. One set may be removed without losing the information contained in the signal.

8.3.5 IF Processing

From modulator M1, the signal is often looped out for possible external signal processing. (The term *loop out* is applied to the routing of a signal from one piece of equipment to another and then back into the first piece for further processing.) Many common scrambling systems (described in Chapter 21) need access to the modulated signal at this point. They attenuate the picture carrier during the horizontal and, usually, the vertical blanking intervals by attenuating the IF signal. Though some older systems attenuated the sync by applying a sine wave attenuation, more modern systems use so-called gated sync suppression, achieved by switching an attenuator into the signal path. The switching operation introduces high-frequency modulation components, which will interfere with adjacent

channel signals if not eliminated. By arranging FL2 so that it follows the addition of the scrambling signal, the undesired modulation components are removed.

After scrambling (when used), the signal returns to the modulator and is supplied to vestigial sideband filter FL2. Vestigial sideband modulation and filtering were described in Chapter 2, and the precise shape of a VSB filter for NTSC is described later. It is used to eliminate one set of modulation sidebands. Modulation sidebands more than 0.75 MHz from the picture carrier are removed from one side of the picture carrier. If the signal has been scrambled using gated sync suppression scrambling, then FL2 will also filter those modulation sidebands.

After filtering, the IF signal is combined with the sound carrier signal. The combined signal may be looped out, at the composite IF loop, to provide for processing by an older sine wave sync suppression scrambling system. It is likely that the signal will be split, with a portion supplied to the composite IF out on the rear panel of the modulator. This signal is used with other processors or modulators in IF switching applications, as described earlier.

The composite (visual and aural) IF signal is up-converted to the RF channel in mixer M3, driven by local oscillator LO3. Attenuator AT5 controls the output level of the composite visual and aural signals. Remaining processing is identical to that described in the signal processor section of this chapter. The signal is filtered, amplified, and filtered again. A phase locking system can be used to phase lock the carrier to another source, and is described later.

8.3.6 Aural Modulation

Figure 8.5 illustrates the modulation process. Baseband audio is supplied to an optional switch S2. If video switching (S1 of Figure 8.4) is used, then audio switching is likely required, too. Normally, the audio switch will be driven by the video switch: when the video switch activates, it activates the audio switch. Sometimes a third audio input, intended for emergency override, is provided though it is not shown in the Figure. Frequency response of the override input may be limited to voice band only. The reason it would be limited is to reduce any noise that may appear on an audio source that is outside the roughly 300–3,000-Hz band needed for understanding voice communication.

Preemphasis

Preemphasis is employed, as in FM radio broadcasting, to reduce the noise introduced by transmission. An FM demodulated signal can be shown to have rising noise density with frequency. In order to compensate, preemphasis is used. Above a certain frequency, the gain in the modulator increases 6 dB per octave (doubling of frequency). The demodulator employs deemphasis, which attenuates the signal by the same amount. In the process, the noise spectrum is flattened, and the total noise heard by the listener is reduced.

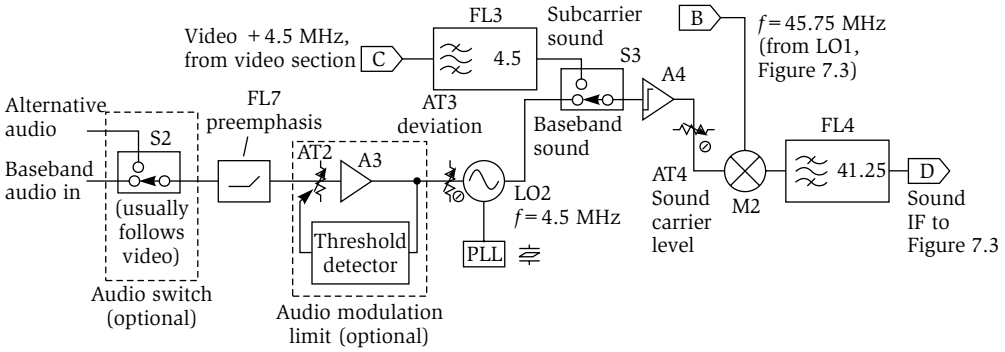


Figure 8.5 Modulator block diagram, aural.

In North America, the standard emphasis curve followed is called the *75-microsecond* emphasis curve because the break frequency, at which emphasis is 3 dB, is given by a 75-μs RC break:

$$\begin{aligned} e_0 &= (1 + j\omega RC)e_i \\ &= [1 + j\omega(75\mu s)]e_i \end{aligned} \tag{8.2}$$

where

$$j = \text{imaginary operator}$$
$$\omega = \text{angular frequency } (\omega = 2\pi f)$$

The corresponding 3-dB frequency, at which the imaginary component is equal to 1, is 2.122 kHz. In Europe and elsewhere, it is common practice to use an emphasis time constant of 50μs, corresponding to a 3-dB break frequency of 3.183 kHz. This affords slightly less reduction in noise, all else being equal, while providing slightly more protection against overmodulation.

The issue of overmodulation with constant amplitude signals arises in the following manner. Consider that you are normalizing the audio level of a program using any of several accepted criteria. You should not set program level based on a preemphasized signal because that is not the way a listener hears the program. On the other hand, the program is transmitted with preemphasis. If the dominant program material occurs at a frequency significantly above the preemphasis break frequency, then overmodulation is possible with normal program level audio. Fortunately, most program material tends to be dominated by frequencies below either the 75-μs or the 50-μs break frequencies. However, sometimes this is not the case, and problems can develop. The 50-μs break frequency provides somewhat more immunity. In order to compensate for the

reduction in signal-to-noise ratio caused by the higher emphasis break frequency, countries using 50-μs preemphasis normally employ wider deviation of the sound carrier (50 kHz compared to 25 kHz for 75-μs countries).

When stereo audio is used, the preemphasis is more complex. The preemphasis circuit in the audio baseband circuits shown is disabled so that the input is flat. Stereo audio is covered in Chapters 2 and Section 8.6.

Aural Modulation Limit, or AGC

An optional modulation limiter may be provided. This circuit monitors the audio signal level, and if the level increases such that the deviation would be excessive, then the modulation limiter temporarily turns down the gain of the signal path by adjusting AT2 of Figure 8.5. After the overmodulation ceases to exist, the gain is returned to its previous value over some length of time, which reduces the audible impact of the change.

Some modern modulators include a second mode of operation, in which the audio level is measured and AT2 is adjusted to provide for a consistent sound level regardless of the incoming modulation. This mode would not be appropriate for a number of audio sources since it will make low-volume sounds the same level as high-volume sounds. (A whisper and a shout would be the same loudness.) It might be appropriate, though, if the program audio consists of normal speech only.

Some modern modulators also include audio clipping, in which audio that would result in significant overdeviation has its peaks clipped, eliminating the overdeviation. In the United States, the FCC requires that a television aural transmitter be able to deviate to at least ±40 kHz, so the appropriate clip point would be above this. The clipping process produces considerable distortion, so operation into the clipping region should never be done intentionally. However, it is a useful way to avoid gross overdeviation, which could cause interference into video or, worse, problems with scrambling recovery (see Chapter 21). It may be applied at all times, so long as the audio level is carefully set to avoid hitting the clipping level during normal programming.

Modulation Process

Audio is frequency modulated onto an aural carrier. Frequency modulation is usually accomplished by applying the audio directly to an oscillator, such that the frequency of oscillation depends in part on the instantaneous modulating voltage. As shown in Figure 8.5, the modulation is applied to LO2, a 4.5-MHz oscillator (other frequencies may be used with conversion).

Some method must be employed to stabilize the center frequency of the oscillator. Though not the only means to accomplish this, a commonly used stabilization method is a phase locked loop (PLL). It generates a frequency correction to the oscillator if the center frequency wanders. The bandwidth of

the PLL is very low to avoid having the loop "Fight" the modulation, which is intentionally forcing the frequency of LO2 to deviate instantaneously.

Audio from Other Sources

If audio is supplied already modulated on a 4.5-MHz subcarrier, then it is introduced into the audio section via flag C (see also Figure 8.4). At this point, the aural subcarrier may be combined with the baseband video, so bandpass filter FL3 is used to separate out only the sound subcarrier. It is passed to one input of switch S3, which selects either the audio from the modulated oscillator, LO2, or the external signal. Switch S3 may be actuated permanently by manual selection of the audio source, or it may be activated by some other event.

A common situation is to activate the switch upon loss of main video. Suppose a modulator ordinarily handles a signal delivered at baseband from a television station. The baseband link may be backed up by an RF link that picks up the off-air signal and demodulates it. The primary audio is delivered in baseband format, and so is applied to the main input to switch S2. The backup demodulator supplies sound on a 4.5-MHz subcarrier, combined with the video. If the video switch (S1 of Figure 8.4) detects loss of video, it activates to couple in alternative video from the backup demodulator. At the same time, it drives the 4.5-MHz switch (S3 of Figure 8.5) to select the external aural subcarrier of flag C.

Following the switch is a limiter whose purpose is to ensure that the sound carrier level is constant regardless of the source. After the limiter is a manual sound carrier level adjustment, AT4, which sets the level of the sound carrier relative to the visual carrier level. Finally, the aural carrier is up-converted from 4.5 MHz to the sound IF at 41.25 MHz using the same local oscillator as was used to generate the picture IF (LO1 of Figure 8.4). The aural IF is coupled at flag D back to Figure 8.4, where it is combined with the visual IF signal.

8.3.7 Vestigial Sideband Filtering

As described earlier, the majority of one sideband is removed by a device known as a vestigial sideband (VSB) filter (FL2 of Figure 8.4). Figure 8.6 illustrates the operation of the vestigial sideband filter in an NTSC modulator. Recall that the spectrum is shown at IF, intermediate frequency, where the picture carrier is higher in frequency than the sound carrier. This is reversed at RF, by the use of a high-side local oscillator.

The VSB filter has a response that is flat from 0.75 MHz above the picture carrier to at least 4.2 MHz below. The reason for the cutoff at 4.2 (really 4.18) MHz is that this is the maximum baseband video frequency that is handled in the NTSC transmitted signal. The color carrier and all its sidebands are contained within the passband of the VSB filter.

The high (vestigial) side cutoff is rather carefully defined. The response should be flat to 0.75 MHz above the picture carrier, then drop to zero at the

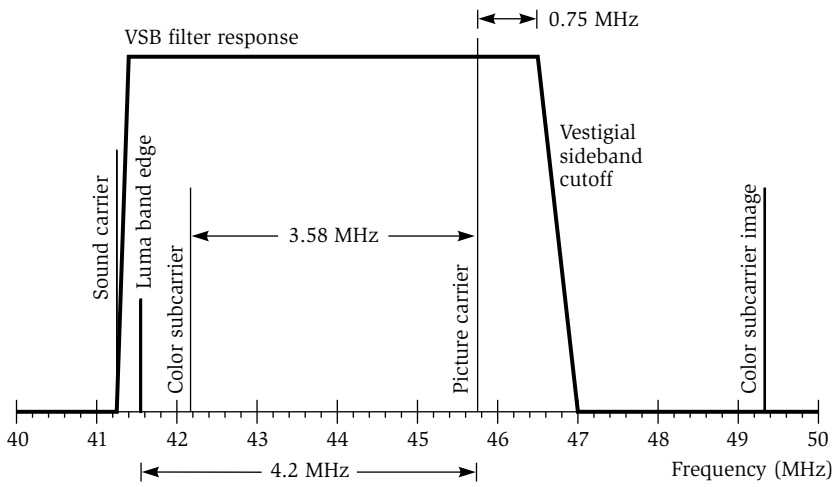


Figure 8.6 Modulator vestigial sideband filter response.

band edge, 0.5 MHz higher. This frequency, 1.25 MHz above the picture carrier, is at the channel edge, 47 MHz at IF.

No filter has truly zero stopband response, so it must be decided how much attenuation is "enough" at the band edge and frequencies farther out. In the section on processors, we argued that the power spilling into an adjacent channel should be at least 60 dB down from the picture carrier. This number is applied to the VSB filter also. The requirement does not mean that the filter must exhibit a stopband response of −60 dB because the power of the modulated TV waveform is not constant throughout the channel. Most of the power is concentrated at the picture carrier. Typically, other carriers are −15 dB or lower with respect to the picture carrier. The only frequency removed by the VSB filter that is likely to be that high is the color subcarrier image at about 49.3 MHz. Thus, if the VSB filter has 45-dB rejection, that is good enough. Even if the filter has only 40-dB rejection, that may be adequate given a small attenuation by other filters.

In some modulators, the sound signal may be supplied through the VSB filter, in which case a slightly wider response than shown is needed. However, with the modulator architecture of Figures 8.4 and 8.5, this would not be necessary.

The critical requirements for the VSB filter made it an early candidate for use of a surface acoustic wave (SAW) device, and in fact, this is the first known application of SAW devices in the cable television industry. SAW devices exhibit sharp cutoff regions, good temperature stability, and low group delay. However, they do exhibit higher propagation delay than do conventional L–C filters. Propagation delay is the absolute delay of a signal as it propagates through the

device. The propagation delay is related to the construction of the SAW device and has decreased somewhat with improvements in the state of the art. All else being equal, the sharper the skirts (the narrower the rolloff region), the greater will be the propagation delay. Delay times of 3 to 5 microseconds tend to be typical of SAW filters built for this application. The propagation delay can be an issue when events must be synchronized between those transmitted on the picture and sound carriers. Such a situation develops in the commonly used scrambling systems where synchronization data is transmitted on the sound carrier (see Chapter 21).

A SAW filter utilizes the propagation of an RF signal on the surface of a suitable material, such as a quartz or lithium niobate crystal. A pair of transducers is built onto the surface of the crystal, using the same techniques as are used to deposit metalization on the surface of integrated circuits. One transducer launches the signal into the crystal, where it propagates as a mechanical wave on the surface. The other transducer converts the wave back to an electrical signal. The shape of the transducers determines the passband of the filter. The loss of a SAW filter tends to be rather high, but the skirts can be made very steep without incurring the group delay that accompanies a sharp L–C filter.

Digital Modulators

Digital modulation is covered in Chapter 4. In general, digital modulators are simpler than their analog cousins. Interfaces with digital modulators are shown later. One issue of concern is when scrambling is added in a digital transmission stream. We use *scrambling* in the cable TV sense of the word, "to render the signal unintelligible to those who are not authorized to receive the signal." In some headend systems, there is an external scrambler that adds scrambling prior to the modulator. In other headend systems, the scrambling is an integral part of the modulator. In some cases, the modulator includes the up-converter to the final channel assignment; in other cases, the up-converter will be external to the modulator.

8.4 Phase Locking of Carriers

The output carriers of both processors and modulators may be phase locked to external sources. Two primary reasons exist for phase locking the output. If the frequency is used for local off-air broadcasting, then there is a high probability that off-air signals will leak into some television receivers or, much less likely, into the cable plant itself. This leakage, or *ingress*, will produce a visible moving pattern on the screen if the on-cable picture carrier is not exactly equal to the ingress frequency. The industry commonly calls this pattern a *beat*. The processor or modulator output can be phase locked to the off-air signal to reduce the visibility of the ingress. This works because the phase locking process forces the frequency of the signal on the cable to be exactly equal to that of the ingress.

The other application for phase locking is to lock all picture carriers on the cable to a master reference frequency. Such techniques can be shown to reduce the interference caused by distortion in the distribution plant. Two techniques have seen fairly widespread use in North America. They are the harmonically related carrier (HRC) technique, in which all picture carriers are locked to harmonics of a master reference oscillator, and the incrementally related carrier (IRC), where all frequencies supplied from the master generator are offset 1.25 MHz. Neither HRC nor IRC techniques tend to be used as much today as in the past, but they will still be found in some locations. Chapter 9 describes the standard HRC and IRC frequency allocation plans.

Yet a third possible requirement for phase locking outputs is described in Chapter 9, in the discussion of node-specific programming. More than one program may exist in a headend on the same channel, and if leakage develops from one to the other, phase locking the two picture carriers will lessen the visibility of the beat. This is true if the subject program is analog. If digital, phase locking is probably not necessary.

The term *phase lock* refers to a technique of forcing one RF carrier to maintain a constant phase relationship with a second carrier. The second carrier is called the *reference carrier* because it is a reference to which the first is locked. If the two carriers have a constant phase relationship to each other, it follows that they are identically at the same frequency. Indeed, a common technique for verifying a phase locked condition is to display one carrier on an oscilloscope that is triggered from the other. The displayed waveform should appear stable on the screen. Note that it is not necessary that the two carriers have a particular phase relationship to each other (though some applications require this), and in general, the phase relation might drift over time.

8.4.1 Phase Locked Circuits and Agile Outputs for Processors and Modulators

Figure 8.7 illustrates one possible block diagram of a phase locked circuit. This can be used with either Figure 8.1 (processor) or Figure 8.4 (modulator). In either case, the output local oscillator, LO3, is replaced by the circuit of Figure 8.7. Mixer M3 is the mixer M3 of Figures 8.1 and 8.4.

The reference frequency, a sample either of the off-air signal or from the comb generator (for HRC and IRC applications) is selected in bandpass filter FL10. It is mixed to the common intermediate frequency of 45.75 MHz in mixer M5. The local oscillator used in this mixing process must be LO3, the same local oscillator used to mix the output channel.

After conversion to IF, the signal is filtered in FL12 before being supplied to limiter A11. This limiter must be used because, if the reference source is an off-air signal, it will have modulation on it, and the modulation must not interfere with the mixing process. From the limiter, the signal becomes one input to phase detector M4. The other input to M4 is a sample of the IF signal developed in the

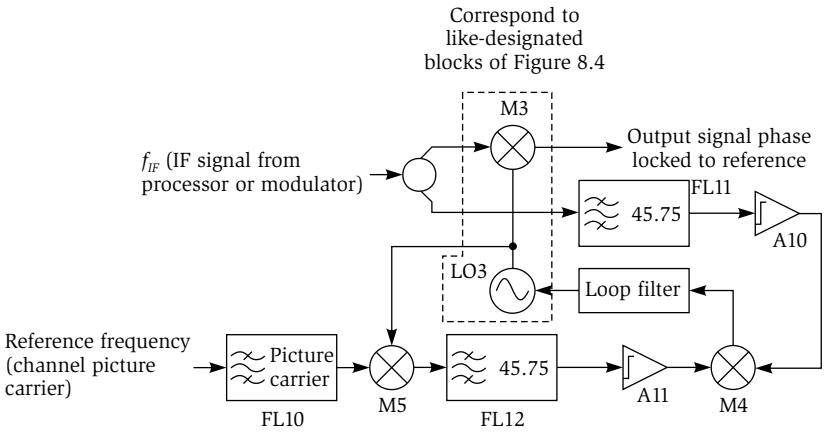


Figure 8.7 Phase lock option for processor or modulator.

modulator or processor. This signal has been filtered (if needed) and limited in FL11 and A10.

Again, a multiplier is shown as the phase detector. A multiplier makes a good phase detector, just as it makes an ideal modulator or mixer. If the reference signal (at frequency ω_r) and the variable signal (at ω_v) are multiplied, the result is expressed as follows:

$$\begin{aligned} E_0 &= \sin \omega_r t \sin(\omega_v t + \theta) \\ &= \frac{1}{2}\cos 2\omega t - \frac{1}{2}\cos \theta \end{aligned} \tag{8.3}$$

where

$$\omega = \omega_r = \omega_v$$

This expansion uses the relation of Equation (2.5). The first term in the second line represents the second harmonic of the carrier frequency and is removed by the loop filter. The second term represents the phase difference between the two carriers, θ.

The output of the phase detector is supplied to the loop filter, which rejects the second harmonic and performs required filtering on the error voltage representing θ. These functions often include providing extremely high gain at dc, which forces $\cos\theta$ to be near zero. The phase locked loop will do whatever is necessary to the frequency of LO3 to keep θ constant. Note that, in order for the cosine of the phase angle to go to zero, the two signals at the mixer will be held

in quadrature, that is, with a 90° phase difference. This phase relationship is typical of multiplicative phase detectors.

If it is not intuitive that the output frequency is phase locked to (that is, the same frequency as) the reference, you are encouraged to write equations for the output frequency in terms of the frequency of the IF, of LO3, and of the reference frequency. Assume that the frequencies of the two inputs to M4 are equal (a condition forced by the phase locked loop), and solve for the output frequency in terms of the reference frequency.

Some modern agile modulators offer an alternative way to achieve phase lock between channels. If all relevant frequencies in all modulators are synthesized from the same reference frequency, then phase lock is accomplished.

Comb Generation Technique

Though not used extensively today, it is instructive to consider briefly how a comb of frequencies might be generated. The comb is used for phase locking processors and modulators in HRC and IRC frequency plans. Figure 8.8 illustrates an HRC comb generator in part a and the modification to make it into an IRC generator in part b.

HRC Comb Generation

The HRC comb is simply all the useful harmonics of a 6-MHz (for NTSC) generator. A simple way to produce the comb is illustrated in Figure 8.8(a). A precision 6.0003-MHz oscillator is used to generate the basic frequency. (The nominal frequency is 6.000 MHz. See Chapter 9 for an explanation of why an oscillator at 6.0003 MHz is used.) After amplification, the signal is supplied to a fast recovery diode and to an inductor. When the signal from A10 goes positive, D10 conducts, setting up current in L10. When the signal from A10 begins to go negative, D10 becomes back biased and cannot conduct. The input impedance to A11 should be high so that no significant current exists in its input. When D10 is back biased, a current path to complete the circuit through L10 does not exist. If a closed circuit does not exist, then the current through L10 must stop. However, if you try to instantaneously interrupt the current in an inductor, the voltage across the inductor must go to infinity (less in a practical case, due to losses in the circuit). This produces a voltage spike that approximates an impulse. An impulse may be shown to contain all harmonics of the fundamental frequency at equal amplitudes. After amplification in A11, the resultant may be applied to filters that remove unneeded harmonics. The small spectrum diagram near the output illustrates the comb.

IRC Comb

Figure 8.8(b) illustrates how the IRC comb is developed from the HRC comb of 6.000 MHz. The IRC comb is the HRC comb except offset up in frequency by

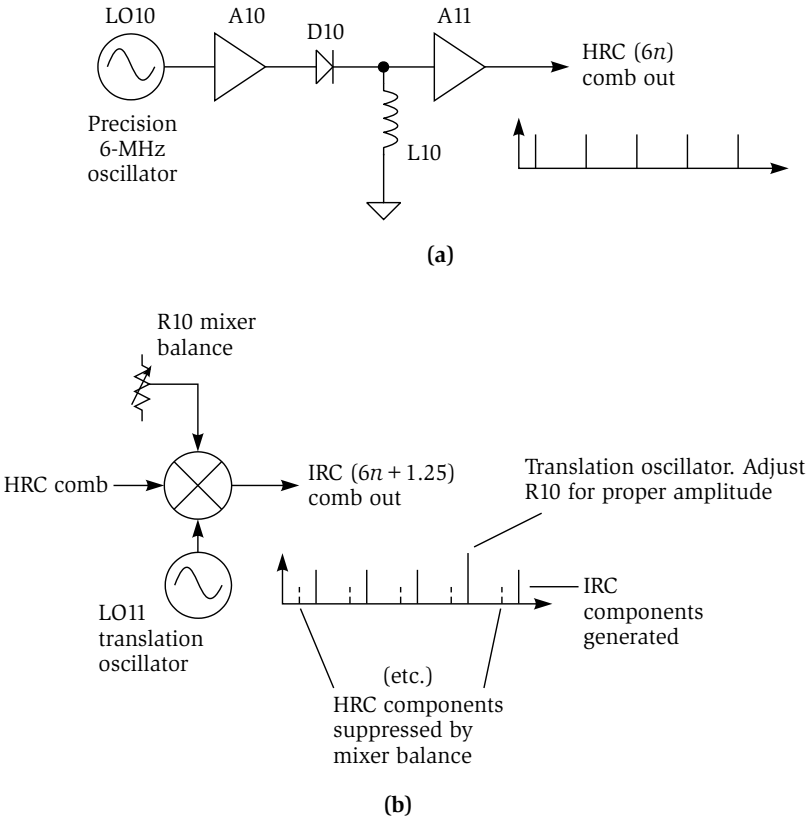


Figure 8.8 Comb generator block diagrams. (a) HRC comb generator. (b) IRC comb generator.

1.25 MHz. This is done by band limiting the HRC comb and using it as modulation on a carrier positioned at a picture carrier frequency of about one-half the maximum frequency. The effect is to generate a carrier at the center of the comb spectrum at one of the comb frequencies. The modulation sidebands are at the other frequencies required of the HRC comb. A double balanced mixer is used for the translation. Since it will try to suppress the frequency of LO11, a mixer balance adjustment is used, and the mixer is intentionally unbalanced to the translation oscillator, enough to cause the translation carrier signal to bleed through at the same level as the sidebands. The HRC comb elements should be completely suppressed, but if they are suppressed by only a few decibels, the phase locked loop in the processor or modulator should not have any trouble recovering the correct element.

Dual Conversion

Figure 8.9 illustrates two features of modern headend processing. One is the use of an agile up-converter to allow the output of a modulator (analog or digital) or CMTS to be placed on any frequency. This generally requires a dual-conversion up-converter to provide low spurious outputs. The genesis of the spurious-output problem is with the output filter after the final conversion. A simple example will show why this is so. Refer to Figure 8.4 for a moment; then we shall return to Figure 8.9.

In Figure 8.4, suppose it was desired to make the modulator frequency agile so that the output could be set to any channel, maybe within a range of several hundred megahertz. We can certainly operate the output local oscillator, LO3, on any frequency (using a phase locked loop to control its frequency), and this will allow us to set the output frequency to any frequency we want. The problem is that we have filters FL5 and FL6, which are tuned to a single output frequency. We will have to either eliminate these filters or make them electrically tunable if we are to allow the output to be on any frequency over a wide range.

But there were some very good reasons as to why we have these filters and why they are only one channel wide. Consider that we want to put the output on channel 13, with the picture carrier at 211.25 MHz (see Appendix A). With a picture IF of 45.75 MHz, this means that the frequency of LO3 will be $45.75 + 211.25 = 257$ MHz. Mixer M3 produces the difference frequency between LO3 and the IF, 211.25 MHz. But the mixer is quite happy also to produce the sum frequency, $257 + 45.75 = 302.75$ MHz, the image of the conversion. We can't let this frequency out of the modulator because it will interfere with channel 37 (and the sound carrier with channel 36). That is one function of filters FL5 and FL6 in Figure 8.4.

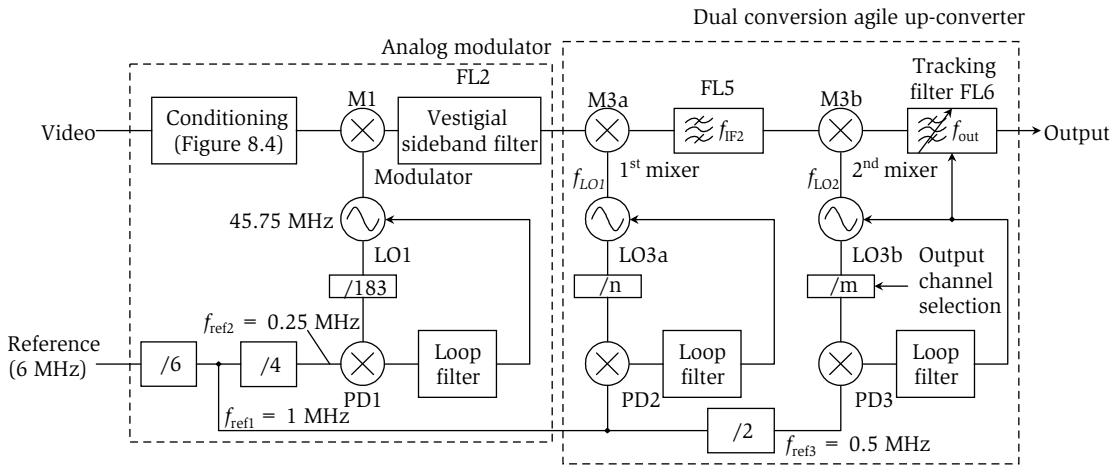


Figure 8.9 Dual-conversion output with optional phase lock to a 6-MHz reference.

So if we want an agile output, we have to modify the block diagram of Figure 8.4. We can use a dual-conversion system such as in Figure 8.9. Like components in Figures 8.4 and 8.9 have the same reference designations, so you can see how they perform similar functions. Video is supplied to all the circuitry shown at the top of Figure 8.9, which we have referred to here simply as *Conditioning*. The actual modulation is done in M1, just as in Figure 8.4. M1 is driven by the IF signal from local oscillator LO1. For phase lock purposes in Figure 8.9, LO1 will need to be phase locked; but it can be simply crystal controlled if phase locking is not to be done (or if it is to be done using the method shown earlier).

The IF signal passes through vestigial sideband filter FL2, just as in Figure 8.4. (Note that this diagram might apply to a digital modulator or to a CMTS, in which case the filter function is still there but is not a vestigial sideband shape and would not be called that.) For simplicity, we'll omit the sound carrier and the IF loop, though they are present. Now what we are going to do that is different from Figure 8.4 is to pass the signal to a dual-conversion agile up-converter. This one is phase locked to an external 6-MHz source, although that is not necessary. The two local oscillators involved in the dual-conversion process will probably be phase locked to a reference, but the reference may be internal to the up-converter rather than external as shown.

The first local oscillator will up-convert the IF to some high frequency. For illustration, assume it converts to a second IF of 927.25 MHz, a somewhat arbitrary frequency. The first local oscillator, LO3a, thus operates at $f_{LO1} = 45.75 + 927.25 = 973$ MHz. The second IF filter, FL5, is set to this frequency and has a width of at least one channel. Its main job is to remove the image of the first IF, $973 + 45.75 = 1018.75$ MHz. We don't show amplification, but there may be some amplification at this second IF frequency before the signal is converted to the final output frequency in mixer M3b, driven by local oscillator LO3b. If we want the output frequency to be 211.25 MHz for channel 13, the frequency of LO3b will be $927.25 + 211.25 = 1138.5$ MHz. Notice that there is still an image of the second IF, but the image frequency is $1138.5 + 927.25 = 2065.75$ MHz, well above any frequency we are interested in on the cable plant, and so far from the output frequency that it is easy to filter.

There is a need for a final filter, however. We show a tracking filter at FL6, which is located after any amplification we provide at the output (amplification not shown). The reason is that there is some noise generated in the mixing process and in amplifiers. If we don't filter this noise, it will be combined with all the other signals and could harm the output C/N of the headend. Thus, a tracking filter can be provided to eliminate this noise. As a practical matter, no tracking filter can simultaneously provide excellent adjacent channel attenuation and flat response in-band while being tuned over a wide range. Thus, some compromises must be made in its ability to attenuate noise in close-in channels versus the flatness and tunability of the filter.

We have shown an alternative way to provide phase lock in Figure 8.9. This uses a reference, in this case 6 MHz, to phase lock all oscillators. The 6-MHz

reference is divided, in frequency dividers, to frequencies that allow it to serve as a reference for all three oscillators. If phase locking is not needed, then there is no need to phase lock the first IF oscillator, LO1. Local oscillator LO3a can either be phase locked or crystal controlled, at the manufacturer's discretion. Local oscillator LO3b will typically be phase locked in any case, since it must be changed in frequency as you set the up-converter to different channels.

Yet another consideration with phase locked oscillators is phase noise produced by the oscillator. For several reasons, the phase noise is typically much greater from a phase locked oscillator than it is from a crystal-controlled oscillator. Phase noise can hurt the bit error rate of a digital signal, as explained in Chapter 4. For analog signals, phase noise can be converted to amplitude noise in the receiver's Nyquist slope filter and can damage the signal-to-noise ratio of the demodulated signal. Considerable effort and expense go into minimizing the phase noise of oscillators in agile up-converters.

8.5 Demodulation

A third type of RF product used in the headend is a demodulator. It is useful to understand the demodulation process in order to make sense of a lot of phenomena observed in cable systems. Also, demodulation is used for quality checks, for recovery of off-air signals for backup purposes, to put signals on a digital transportation route, and as interfaces for FM microwave systems. A few headends employ demodulation and remodulation instead of processing for off-air signals. This is advantageous if considerable signal switching is done. The use of a quality demodulator as a test tool is, alone, justification for studying the technique.

8.5.1 RF Input Section

Figure 8.10 shows the RF, IF, and video demodulation portions of a demodulator. The sound section is shown in upcoming Figure 8.11. The demodulator input is identical to that of the processor of Figure 8.1. Both processor and demodulator are used to recover signals from off the air or from an incoming cable. Thus, the input circuitry through attenuator AT2 is identical.

8.5.2 IF Processing Section

IF processing in a demodulator is significantly different from that of a processor. Power for demodulation of the sound carrier is taken at flag A. Sound demodulation will be covered later in conjunction with Figure 8.11. The AGC operates to hold the visual signal level after AT2 at a constant level regardless of the input level.

Zero chop switch S1 is optional but is often provided on professional demodulators as a way to measure visual signal depth of modulation. When

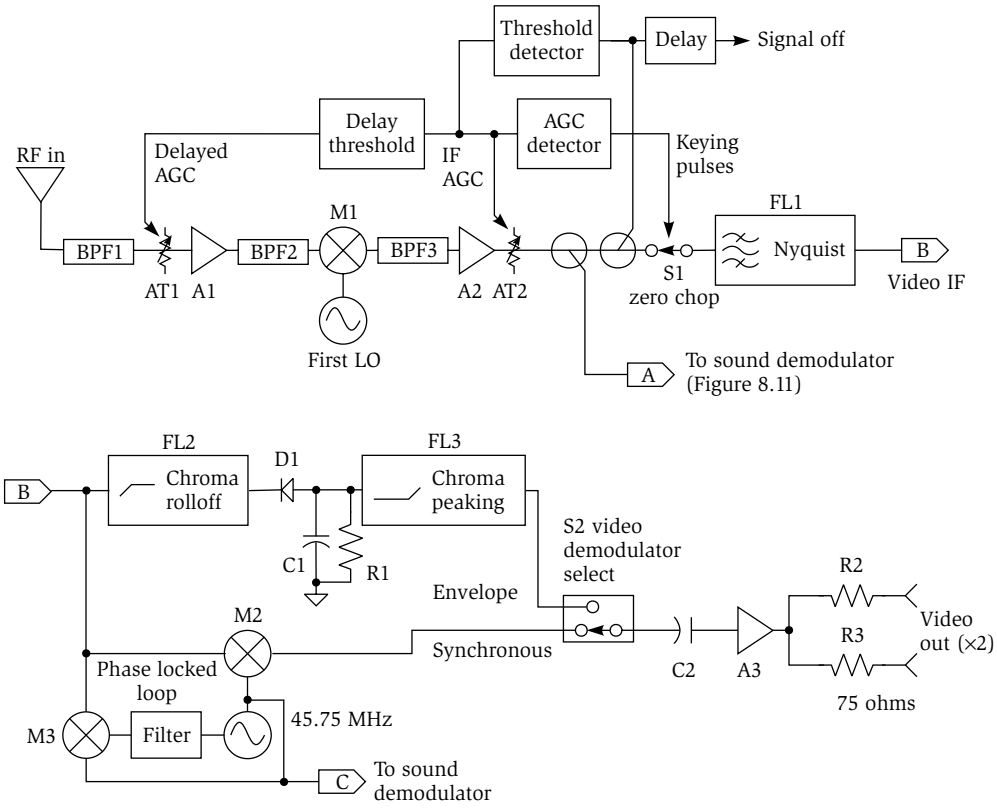


Figure 8.10 Demodulator block diagram, input and video.

enabled, the chopper momentarily interrupts the signal at IF, usually during the vertical blanking interval. The interruption allows a waveform monitor or oscilloscope to be calibrated to measure visual depth of modulation very accurately. Chapter 9 describes measuring depth of modulation using a zero chopper. The chopper is coupled to the AGC detector in the illustrated case so that the switch may be restored to the IF path during sync pulses. By doing this, the zero chop pulse may extend over several lines without interrupting sync.

From the zero chop circuit, the signal is supplied to the Nyquist slope filter, FL1, which provides the main filtering function for the visual signal. The filter is named after Harry Nyquist, a Bell Laboratories engineer who developed this technique for recovering vestigial sideband signals. The filter linearly reduces the amplitude response in the double sideband region (picture carrier ± 750 kHz) such that the picture carrier is reduced 6 dB. The reason for doing this is to equalize the frequency response in the double sideband and single sideband regions.

8.5.3 Visual Carrier Demodulation

After the Nyquist slope filter, the signal is supplied to one of two demodulation circuits, an envelope detector or a synchronous detector. Home television receivers use one detector or the other, but it is advantageous to have both in a professional demodulator because the two will yield different information about a signal. Some of the differences are discussed in Section 8.5.7.

Envelope Detector

In the illustrated system, the envelope detector consists of a chroma rolloff filter, FL2, and following circuits. The chroma rolloff filter reduces the amplitude of the IF signal at the color subcarrier frequency. The purpose of the chroma rolloff filter is explained later. Diode D1 performs the actual detection process, and capacitor C1 filters the signal. The detector, except for component values, looks just like the detector for an AM broadcast receiver.

The direction of the diode is not arbitrary. Recall from Chapter 2 that the sync tip is the most negative portion of the video signal as the signal is normally interfaced. The sync tip also corresponds to the highest amplitude of the signal. With the diode in the direction shown, the voltage across C1 will be such that the highest amplitude of the RF signal (the sync tip) will be the most negative voltage across C1. Other voltages will be positive with respect to that. Of course, in a real demodulator, it is possible to detect the signal with the opposite polarity and then invert the video, but this is often not done for practical reasons.

Following the detector is the chroma peaking filter FL3, which restores flat chroma response, complementary to the operation of the chroma rolloff filter FL2. The signal is then applied to switch S2, which selects either the envelope detector or the synchronous detector.

Synchronous Detector

The synchronous detector is shown below the envelope detector in Figure 8.10. It consists of a phase locked loop used to generate a carrier phase coherent with the incoming picture carrier (at IF). In contrast to other applications described earlier, the phase of the locked carrier with respect to the input IF signal is of paramount importance: the two carrier phases must be the same. Because the phase of the carrier varies with modulation and, in addition, can vary owing to defects in transmission, the phase locked loop is normally keyed (closed) on either sync tips or the back porch portion of the modulated signal.

When a modulated carrier is multiplied with an unmodulated carrier at the same frequency, the result is recovery of the baseband signal. The advantage of the synchronous detector over the envelope detector is primarily associated with demodulation of signals having asymmetrical sidebands, which is the case for all

analog television signals. An envelope detector introduces distortion that is avoided with a synchronous detector.

From the synchronous demodulator M2, the signal is low-pass filtered to remove the carrier and harmonics, and supplied to S2, the switch that selects either the envelope or synchronous detectors.

8.5.4 Video Output

Switch S2 is normally actuated manually to select either the envelope or synchronous detectors if both are provided. By comparing the results of demodulation using the synchronous and envelope detectors, we can deduce a lot about the performance of the RF components handling the signal. Under normal conditions, the synchronous detector will yield a more accurately demodulated signal than will the envelope detector.

A coupling capacitor C2 removes dc components of the demodulation process, as well as any bias voltages introduced in the demodulation process. It has been customary to ac couple the video output since this minimizes the power supplied or dissipated in the load and often simplifies circuitry. Today, both dc and ac coupling of the output are employed. If dc coupling is employed, capacitor C2 would likely be used anyway, with a dc restoration circuit in the output amplifier.

It is essential with video that the coaxial cable impedance be matched both on the source end and the load end. Failure to do so can cause unacceptable frequency response errors, due to reflections, if the interconnecting cable is long enough. A typical way to achieve multiple well-matched outputs is to provide a very low output impedance in amplifier A3 and multiple matching resistors R1 and R2 at the outputs.

8.5.5 Aural Carrier Demodulation

The aural demodulation process is shown in Figure 8.11. The sound IF is supplied from Figure 8.10 on flag A. The other input to the sound demodulation

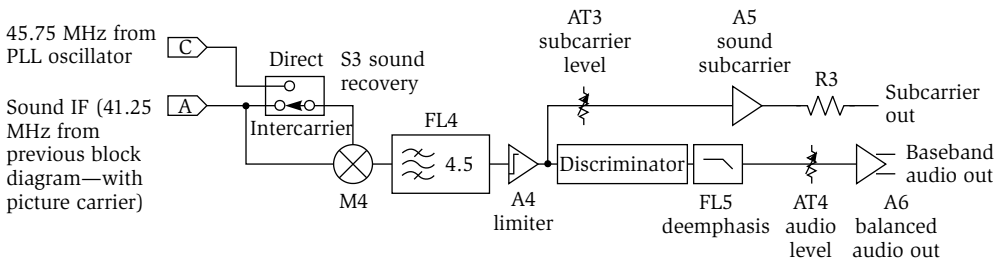


Figure 8.11 Demodulator block diagram, audio.

process is a sample of the signal from the 45.75-MHz phase locked oscillator, used in the synchronous demodulator (flag C, which is omitted if a synchronous detector is not supplied).

The intercarrier detection mode is the one used almost exclusively in consumer applications. The sound final IF frequency, normally 4.5 MHz, is derived from mixing the sound first IF, 41.25 MHz, with the picture carrier, 45.75 MHz. This final sound IF frequency is often called the *intercarrier frequency*. Use of intercarrier detection solves several problems related to television transmission. If any process in the transmission of the television signal introduces phase noise, then the phase noise will seriously damage the signal-to-noise ratio of the audio: phase modulation is a form of frequency modulation, so once phase noise is introduced, it cannot be removed. However, since phase noise will affect the picture and sound carriers identically, when the intercarrier signal is developed by taking the difference in the picture and sound carriers, phase noise is eliminated from the aural signal. The IF signal on flag A contains both the aural and picture IF carriers, so by applying this signal to both inputs of a mixer, intercarrier mixing results. (Additional filtering may be used.)

On the other hand, if phase noise is not a problem, then there are advantages to a direct detection process. The picture carrier may have some incidental phase modulation on it, and it certainly has visual modulation sidebands. These will cause interference with the aural signal if used in the intercarrier detection process. In direct detection, the CW signal locked to the picture carrier is mixed with the sound carrier to derive the intercarrier frequency. (In this mode, the sound carrier path may include filtering to remove the picture carrier though this filtering is not shown in Figure 8.11.)

After the intercarrier detector, M4, the signal is filtered in a 4.5-MHz bandpass filter FL4. Following the bandpass filter is a limiter, A4, which removes any residual amplitude modulation from the aural carrier. At this point, a sample of the aural carrier is usually taken. It is passed through a manually controlled attenuator AT3, amplifier A5, and a current source resistor R3. The purpose is to provide a sound subcarrier for external use. Often this signal is combined with the picture carrier to provide a signal to be transmitted by FM microwave link. The resistance of R3 is high compared with 75 ohms. If one of the video outputs on Figure 8.10 is paralleled with the subcarrier out connector, then R3 will act as a current source, adding the 4.5-MHz subcarrier to the video without significantly upsetting the impedance of the video circuit.

The other path from the limiter is through a discriminator, which recovers baseband audio. Note that the signal recovered at this point is either monaural (mono) or a stereo composite signal, as defined in upcoming Section 8.6. If the signal is mono, then a deemphasis filter FL5 is included. If it is desired to recover a composite stereo signal for subsequent demodulation or processing, the deemphasis is omitted in favor of more complex processing in the stereo decoder.

Finally, the signal level is adjusted at attenuator AT4, normally a manual control. It is amplified in output amplifier A6 and provided on balanced outputs.

Balanced outputs are used in professional equipment to provide for reduction of pickup of 60-Hz ac components ("hum"). "Balance" refers to a transmission technique in which the two sides of the output have equal impedance to ground, as opposed to unbalanced transmission, where one side of the circuit is at ground potential. Section 8.9.2 describes use of balanced interconnection for audio.

From a long-standing telephone specification, the impedance of the audio output is normally 600 ohms between the two balanced lines. In much equipment today, the source impedance is less than 600 ohms, but the performance of the amplifier is specified when the amplifier is working into a 600-ohm load. The lack of a defined 600-ohm output impedance is not a problem in most cases since the length of audio cable is very small compared with the wavelength of any audio frequency. If the output is to drive a line several miles long, then the impedance should be controlled more carefully.

8.5.6 Nyquist Slope Filter

The Nyquist slope filter shown as FL1 in Figure 8.10 has a shape shown by the solid line in Figure 8.12. At the low-frequency side (aural carrier at IF), the shape is fairly straightforward. The bandpass characteristic should pass the color carrier and its sidebands while rejecting the sound carrier.

The high-frequency side (near the picture carrier) is rather unusual. The purpose of this shape is to complement the shape of the vestigial sideband (VSB) filter used in the modulator, and shown in Figure 8.12 as a dotted line. The

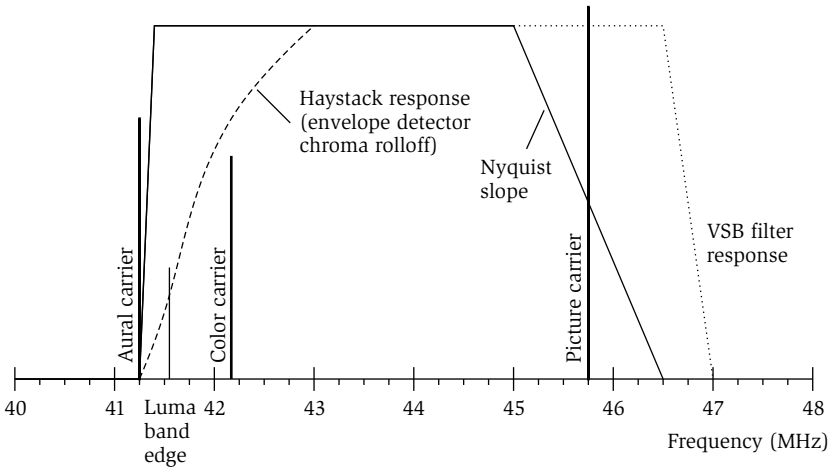


Figure 8.12 Nyquist slope filter response, with haystack.

Nyquist shape is flat to 0.75 MHz below the picture carrier; then it begins falling off at a prescribed rate above that frequency. The rolloff must be linear with frequency on a voltage plot, not a logarithmic plot. The response of the filter must be 0.5 at the picture carrier, continuing to drop until the response is zero at 0.75 MHz above the picture carrier.

Other features of the Nyquist slope filter are emphasized in the next section, in which we describe the differences between synchronous and envelope detection. One of those features is not really a part of the Nyquist slope filter response. It is shown in Figure 8.12 as a dashed line and was mentioned in conjunction with the chroma rolloff filter (FL2 of Figure 8.10). This is the reduction of response near the color carrier applied in the case of an envelope detector. It is called a *haystack* response after traditional television usage: taken with the Nyquist slope around the picture carrier, the filter shape looks like a haystack. The haystack response is used to reduce the effects of luminance-to-chrominance crosstalk.

In consumer television receivers using envelope detectors, the haystack response is built into the Nyquist slope filter (FL1 of Figure 8.10). In professional receivers, the chroma rolloff may be realized independently of the Nyquist slope filter so that it is not in the signal path in the synchronous detector mode. Not that the haystack response would not hurt the response of the synchronous detector if it were perfectly compensated at baseband. Rather, it is difficult to compensate perfectly, and for measurement purposes, it is better not to introduce a response error that must later be compensated.

In modern practice, the Nyquist slope filter is normally realized as a SAW filter in consumer receivers though the shape is compromised to allow the SAW filter to be built with good cost effectivity. In professional demodulators, the filter may be realized either with a SAW device or as a conventional L–C filter with delay equalization.

8.5.7 Synchronous Versus Envelope Detection

Both synchronous and envelope detectors may be used for operational demodulation. Synchronous detection yields the higher-quality results absent a few distortions that prevent its proper operation. The most common problem that interferes with operation is phase noise, or phase modulation, on the picture carrier. A synchronous detector will not work as well as does an envelope detector under this circumstance.

Envelope detectors are susceptible to the practical limitations of diodes though there are ways to mitigate this problem. More seriously, envelope detection is susceptible to quadrature distortion errors that arise as a result of asymmetrical sidebands in the television signal. Quadrature distortion occurs when the upper and lower sidebands are not symmetrical. As shown earlier and in Chapter 2, the television signal presented to the detector does not have symmetrical sidebands at any frequency: at modulation frequencies above 1.25 MHz,

only one sideband is transmitted. Below 0.75 MHz, symmetrical sidebands are transmitted but are turned into asymmetrical sidebands by the Nyquist slope filter. The lowest modulating frequencies approach sideband symmetry, but as the frequency increases, the symmetry is quickly lost.

Phasor Representation of the Modulated Signal

The mathematics of modulation were shown in Chapter 2. A useful graphical representation that shows the function of the two demodulators is to invoke the mechanism of phasors.[2] A phasor represents any steady-state sinusoid as a vector whose length is a function of the magnitude of the signal and whose angle is a function of the angle between a reference and the phasor. From Equation (8.1), an amplitude modulated signal can be represented as

$$\begin{aligned} V_0 &= \cos \omega_c t(1 + M \cos \omega_m t) \\ &= \cos \omega_c t + \frac{M}{2} \cos(\omega_c - \omega_m)t + \frac{M}{2} \cos(\omega_c + \omega_m)t \end{aligned} \tag{8.4}$$

Figure 8.13(a) represents this equation in phasor form. The first term, representing the carrier, is defined as having a phase such that its phasor is straight up (this is arbitrary). The upper sideband has a phasor that rotates clockwise (with respect to the stationary carrier) at an angular frequency of ω_m. The lower sideband term, $\cos(\omega_c - \omega_m)$, is represented by a counterrotating phasor. The resultant phasor is shown to the right in Figure 8.13(a). The vector sum of the three phasors is always in phase with the carrier. An envelope detector responds to this resultant vector. Analysis of the vector at all different phase angles θ would show that the resultant traces out a sinusoid, which is the waveform used to modulate the carrier (the $M \cos \omega_m t$ term of Equation (8.1)). All is well.

Now consider the situation shown in Figure 8.13(b), a single sideband signal. The modulating waveform is the same sine wave as in Figure 8.13(a), but one sideband has been removed by the VSB filter (FL2 of Figure 8.4). Most notably, one sideband of the color subcarrier is removed, so the color subcarrier is transmitted as a single sideband signal. The vector sum of the carrier and single sideband phasor is shown to the right of Figure 8.13(b). Here the resultant is *not* in phase with the carrier. The resultant can be resolved into an in-phase component and a "quadrature" component at right angles to the carrier phase.

Response of an Envelope Detector

The envelope detector responds to the magnitude of the resultant carrier. By applying trigonometric relations to the resultant vector of Figure 8.13(b), we can compute the magnitude of the vector for various conditions. Figure 8.14 illustrates the result obtained using an envelope detector for two different relative levels of

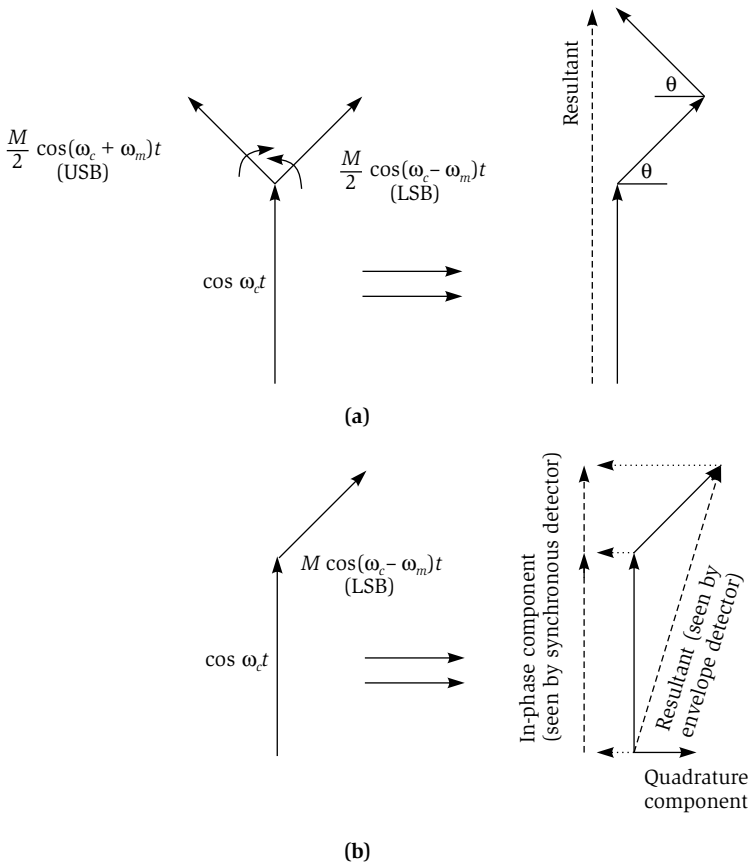


Figure 8.13 Phasor diagram, double and single sideband carriers. (a) Symmetrical sidebands. (b) Single sideband.

carrier and sideband.* The dashed line is the sinusoid with which the carrier was modulated. One of the two cases represented (the middle trace) results when the sideband is −6 dB (one-half the voltage) of the carrier. The third waveform represents the waveform produced by equal amplitude carrier and sideband. This is a realistic condition for a television signal: a scene with medium to high brightness has a relatively low carrier level during the lighter portions of the

* The Figure is displayed as would be a television signal, with the lowest resultant amplitude at the top. The gain has been normalized in the three cases to produce the same peak-to-peak signal.

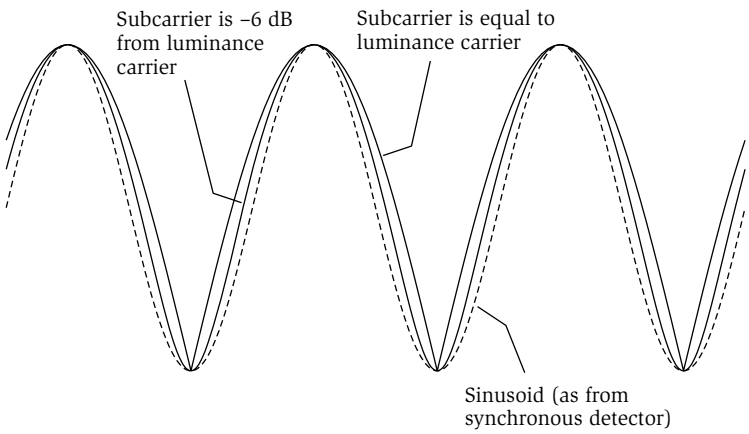

Subcarrier is –6 dB from luminance carrier
Subcarrier is equal to luminance carrier
Sinusoid (as from synchronous detector)

Figure 8.14 Demodulated results, envelope detector.

scene. If the object is brightly colored, the chroma carrier amplitude will be high, so when the camera is scanning the object, the chroma can be as high or higher in amplitude than the picture carrier. The resultant tends to be "scalloped," with the negative peak too wide and the positive peak too narrow.

From this illustration, it is evident why the chroma response of the Nyquist slope filter is reduced (haystacked) when an envelope detector is used (see Figure 8.12). The chroma carrier is the highest amplitude sideband of the luminance carrier when brightly colored objects are being scanned. The higher the chroma signal is with respect to the picture carrier, the greater the distortion. By reducing the chroma amplitude before detection, the effect is reduced. Of course, the chroma amplitude must be corrected before the signal is presented for measurement purposes; hence the chroma peaking circuit FL3 of Figure 8.10. (In a consumer TV set, it is not necessary to use this peaking circuit because the chroma is separated from the luminance, and the correction can be approximated by adding more gain to the chroma channel.)

If the Fourier transform of the chroma waveform is taken, it will be seen that the amplitude of the 3.58-MHz fundamental component is less than what it would have been without quadrature distortion. This causes the chroma saturation (proportional to chroma amplitude) to be lower than it should be — the color appears washed out compared with the way it should look. An envelope detector that doesn't reduce the amplitude of the color carrier prior to demodulation will exhibit excessive differential gain and luminance-to-chrominance and chrominance-to-luminance crosstalk.

The average level of the chroma signal is no longer in the center of the waveform of Figure 8.14. Rather, it is lower than it should be. This means that

the luminance level is lower than it should be, making the object appear darker than it should. Thus, a brightly colored object of relatively high luminance level (bright yellow is an example) will appear too dark and somewhat washed out owing to the quadrature distortion introduced by the envelope detector. Whether or not the distortion is objectionable depends on a number of factors.

One useful way to show the difference between envelope and synchronous detectors is to study the various sinusoids involved in the multiburst waveform. Switching back and forth between the two detectors will readily show the difference in waveform. Some bursts will show more distortion with an envelope detector than will others, depending on the sideband amplitude seen by the detector.

Another convenient test is to study the difference between a sine squared (2T) pulse when demodulated with synchronous and envelope detectors. The demodulated shape will be correct with the synchronous detector. With the envelope detector, the pulse will appear too narrow, and possibly too tall. This would have the effect on a picture of narrowing a vertical white line on the screen.

Response of a Synchronous Detector

The synchronous detector can be shown to respond only to the in-phase component of the phasor of Figure 8.13(b). In turn, it can be shown that the in-phase component, the projection of the sideband on the phase of the carrier, is an accurate representation of the modulation.

Application of Test Demodulators Having Both Detectors

Some demodulators have a second synchronous detector, in which the phase of the reconstructed carrier is 90° with respect to the incoming picture carrier (at IF). This quadrature detector is sensitive to the quadrature component of Figure 8.13(b). By displaying the quadrature component on the *y* axis of an oscilloscope and the in-phase component on the *x* axis, we can easily spot incidental carrier phase modulation (ICPM) in the modulator. Normally, a luminance ramp is supplied as video for this test. ICPM can be caused by improperly operated power amplifiers, and is a common problem with older broadcast transmitters, particularly UHF transmitters. It tends to be less of a problem with cable modulators because of lack of high power stages, but it can be introduced in an amplifier stage that is distorting. ICPM can cause excessive intercarrier buzz if it affects only the luminance carrier.

A problem with some older and lower-cost cable modulators is poor balance in the video modulator stage. This will distort the luminance level of bright portions of a picture. The presence of this may be deduced by measuring differential gain with the envelope and synchronous detectors. A good differential gain reading with the synchronous detector and a bad reading with the envelope detector indicate possible problems with the modulator's modulation stage.

Subtle amplitude and group delay problems in a piece of equipment can be masked when measured with an envelope detector. Amplitude and group delay problems can be evaluated by looking at the sine squared (2T) and modulated 12.5T pulses. These pulses are particularly convenient for use in evaluating problems because the results are presented in terms that directly relate to the visible distortion of a video signal. However, some problems produce waveform distortions that are similar to the distortions produced by an envelope detector. Thus, a synchronous detector is more useful when studying response problems.

8.6 TV Stereo

At least three TV stereo systems are in use worldwide, with several variations to each. North America uses the BTSC system, named after the committee that adopted it. The Broadcast Television System Committee was formed under the auspices of the Electronic Industries Association (EIA). It was active from the late 1970s to the mid-1980s. The Japanese use a variation of BTSC stereo. In Great Britain, the BBC and IBC developed a digital system known as near instantaneous amplitude companding and modulation (NICAM). With variations, it is in use in a number of European countries and elsewhere. The German IRT earlier developed a two-carrier analog system, which is used there, in Korea, and elsewhere.

8.6.1 BTSC Stereo Generation and Application

The BTSC stereo system is described in Chapter 2. Figure 8.15 shows one possible block diagram of a BTSC stereo encoder. Left- and right-channel audio are supplied to identical low-pass filters, whose cutoff frequency should be as close to 15 kHz as possible, while providing a lot of attenuation at the horizontal sweep frequency of 15.734 kHz. Because of the sharpness demanded of these filters, their complexity is high. The filters are often implemented using digital signal processing (DSP) technology today. Failure to provide adequate reduction of power at 15.734 kHz will result in compromised stereo performance due to interference with the pilot carrier. Further, the difference channel approaches this frequency, and poor rejection could result in artifacts arising from spectral crossover between the sum and difference channels.

Note that the left- and right-channel program audio levels must be metered as shown to ensure the correct deviation. The monaural meter on the modulator is not usually suitable for this purpose.* In fact, it can be very misleading. Similarly, SAP level is monitored at its input.

---

* If the monaural meter drive signal is limited to a 15-kHz frequency response it may be marginally useful as a deviation indicator, but this is not recommended.

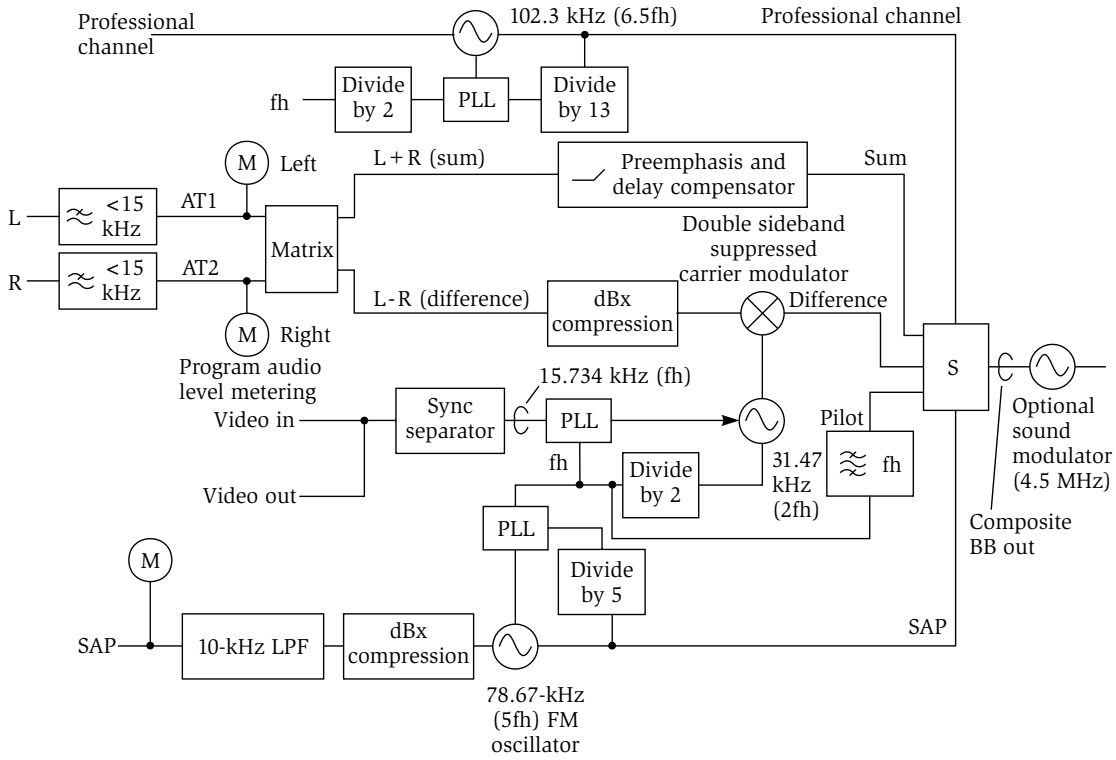


Figure 8.15 Block diagram, BTSC stereo encoder.

From the filters, the signals are supplied to the matrix, which is simply a pair of networks that produce the sum (L + R) and difference (L − R) signals. The sum channel is supplied to a preemphasis circuit and a delay compensator, which compensates for the delay in the difference channel. See the discussion of emphasis in Section 8.3.6. It is necessary to use preemphasis in the sum channel in order to maintain backward compatibility with the mono system. From the preemphasis circuit, the sum signal is supplied to an adder that combines the different parts of the stereo signal.

The difference signal is supplied to the dBx encoder, which compresses it in order to reduce the effects of poor carrier-to-noise ratio in the transmission path. The compression is frequency selective and follows an algorithm proposed originally by the dBx Corporation. (For test purposes only, the dBx compression may be replaced with a 75-μs preemphasis network.)

From the compression circuit, the difference signal is applied to the modulator. It is double sideband suppressed carrier amplitude modulated onto a

carrier at 2fh (twice the horizontal scan frequency), 31.47 kHz. The carrier oscillator is phase locked to the horizontal sync rate, fh, derived from the video signal. This need for phase lock is the reason that video must be looped through the stereo encoder. Failure to lock to video could result in beats and separation "pumping" from oscillators that are free running on almost but not quite the same frequency. A portion of the fh signal, obtained by dividing the output of the 2fh oscillator by 2, is filtered and supplied to the summing circuit as the pilot carrier. This fh signal is also supplied to a phase locked loop, whose output adjusts the 2fh oscillator as needed to maintain the correct phase relationship between its two inputs, one from the divider and the other from the sync separator. These three components — the sum, difference, and pilot signals — make up the basic stereo signal.

If the SAP is not used on a particular channel, that circuitry is not likely to be included in the stereo encoder. If it is, the circuitry at the bottom of Figure 8.15 is supplied. The SAP audio is compressed then frequency modulated onto an oscillator whose center frequency is locked to the fifth harmonic of the horizontal frequency, 5fh. This SAP signal is then summed with the stereo components. If the professional channel is used, then the blocks at the top of the page are supplied. It is rarely used in cable television.

Interfacing and Setting Levels in BTSC Service

A number of methods may be used to connect the BTSC encoder to the modulator. Selection of the interface method to use is based on inputs available on the modulator and outputs available on the stereo encoder. Of particular concern is the ability to set the deviation of the sound carrier correctly. Failure to do so will result in decreased stereo separation. For example, separation between the left and right channels will be degraded to less than 20 dB if there is a 1.9-dB error in setting the deviation.[3]

Several methods may be used to interface the encoder with the modulator. The most basic, but not necessarily the best, is to defeat the preemphasis in the baseband input of the modulator (because preemphasis is handled in the encoder) and supply the broadband signal to the baseband audio input of the modulator (Figure 8.5). This works, but you must be careful to set the deviation of the modulator correctly. Failure to do so will result in poor stereo separation, as shown earlier.

In order to use this interface, some modulator designers have arranged metering circuitry such that a modulation meter measures only the sum signal. The object would be to set the level using the left and right meters on the encoder, then set the deviation on the modulator using its meter. Setting deviation only with reference to the sum deviation is not at all accurate and will almost surely result in poor stereo separation. However, metering the sum channel at least makes a meter on the modulator make sense when handling a stereo signal.

In addition to metering the program audio, the encoder manufacturer may provide a test tone. When turned on, the test tone replaces the composite stereo signal supplied to the modulator. The modulator *deviation* adjustment (AT3 of Figure 8.5) is adjusted for the correct deviation of the test tone, usually 25 kHz. If this is done and the program audio levels are set correctly on the encoder, then the deviation will be correct. This method is far superior to metering only the sum channel program audio on the modulator.

A preferred interface method, if available, is to have a 4.5-MHz FMed oscillator built into the stereo encoder. An optional sound carrier modulator (oscillator) is shown in Figure 8.15. This way, the encoder manufacturer can ensure that the deviation is set correctly. The encoder output at 4.5 MHz is supplied to flag C of the modulator block diagram, Figure 8.5. It is best to *not* combine the sound subcarrier with the video since doing so will reduce the carrier-to-noise ratio in the aural channel and will create more group delay problems in the video channel. The modulator must be capable of accepting a separate input to FL3.

Another possible but somewhat less desirable interface method is to locate a 41.25-MHz FM oscillator in the encoder and supply sound at this frequency to flag D of the sound IF loop at the modulator, Figure 8.4. This provides the same level calibration advantage as does the 4.5-MHz interface, but exacerbates the problem of maintaining accuracy in the frequency spacing between the picture and sound carriers. In any of the other methods described, sound carrier spacing depends only on a single 4.5-MHz oscillator. The FCC-required accuracy in spacing for cable television is ± 5 kHz. Applied to the 4.5-MHz spacing, this represents an accuracy requirement of 1,100 parts per million, a trivial number for a crystal-controlled oscillator, and barely possible with an L–C oscillator (used in set top terminals but definitely not appropriate for professional use). If the interface is at 41.25 MHz, then two independent oscillators are involved, one at 41.25 MHz in the encoder and another at 45.75 MHz in the modulator. This means that each must be accurate to about 57 parts per million. This is not excessive for a crystal oscillator but is getting to the point that care must be exercised.

In addition, there is some possibility with a 41.25-MHz interface that ICPM could exist on the picture carrier and not on the sound carrier. This would cause a degraded signal-to-noise ratio in the intercarrier demodulator in the TV receiver. (See the discussions of ICPM and intercarrier detection in Section 8.5.)

Regardless of the interfacing method used, it is necessary to ensure that the deviation of the main sound carrier by the composite stereo signal is correct, then to set the level of the left and right channels using only attenuators AT1 and AT2 of Figure 8.15. The stereo encoder should provide metering to allow accurate setting of the program audio levels.

8.6.2 The NICAM Stereo System[4]

NICAM's formal designation is NICAM 728 because the bit rate is 728 kb/s. The system is a digital transmission system developed in the mid-1980s in the United

Kingdom. It has been adopted for use in other countries as well. It is capable of carrying one stereo pair, or two independent monaural programs, or 352-kb/s data and one monaural program, or 704-kb/s data only.

Data is initially sampled at 14 bits per sample and down-converted to 10 bits (plus parity). The data is divided into blocks of 32 samples, and the scaling of the block is based on the highest amplitude sample in that block. A scaling factor is transmitted with the data. Transmission is divided into 1,000 frames per second. Each frame consists of 704 bits of data and parity, with the remaining 24 bits being overhead and synchronization.

Modulation is differential QPSK (see Chapter 4). As practiced with the CCIR-I television system (Great Britain), the modulator and demodulator both use a root raised cosine filter with an alpha of 100%. When NICAM is used in countries employing the CCIR-B/G television format, the filter alpha is reduced to 40% to accommodate the narrower available bandwidth.

In the CCIR-I system (used in Great Britain), the NICAM carrier is located 6.552 MHz (nine times the 728-kHz bit rate) above the picture carrier. The normal analog sound carrier, 6 MHz above the picture carrier, is transmitted −7 dB with respect to the picture carrier, and the NICAM signal is set to $-20\ dB_{PC}$. The power in the NICAM signal does spill out of the channel, slightly invading the upper adjacent channel. This has not been a problem.

In the -B/G system (used in much of continental Europe except France and Russia), the NICAM center frequency is slightly outside the channel, with energy spillage even farther out. At UHF (system -G), the occupied bandwidth falls in the 1-MHz guard band between the channels. At VHF (system -B), the power spills into the upper adjacent channel. By narrowing slightly the vestigial sideband filter for the upper adjacent channel, it is claimed that a cable system can successfully add the NICAM signal while employing adjacent channel transmission.[5]

8.6.3 The IRD Two-Carrier Stereo System

An earlier stereo system developed in Germany employed two analog carriers. It is used there with the CCIR-B/G television system. It has also been used, with modification, in Korea, where NTSC is the television system in use. The normal 5.5-MHz (352fh) sound carrier is modulated with a left plus right signal. A second carrier is added at 5.7421875 MHz (367.5fh). It is modulated with the right channel only. This second signal is demodulated and supplied directly to the right loudspeaker. The left signal is obtained by subtracting the right signal from the L + R signal on the main aural carrier. The reason for doing this rather than placing a difference (L − R) signal on the second carrier is that noise from the second, low-amplitude, carrier theoretically gets canceled by virtue of its phase being opposite on the two channels.

The main sound carrier at 5.5 MHz above the picture carrier is carried 13 dB below the picture carrier. The second sound carrier is carried −20 dB from the picture carrier and is also frequency modulated with a deviation of ±50 kHz.

A pilot carrier of 64.6875 kHz (3.5fh) contains modulation that identifies the broadcast as mono, stereo, or dual sound (two mono channels, as for two languages). In addition, mode identification is carried in the vertical blanking interval of the video.

8.7 Satellite Earth Station Receiving Equipment

The modern cable television industry was enabled by the availability of communications satellites that could distribute many television programs simultaneously to all cable headends in North America. The watershed event was the September 30, 1975 (U.S. time), Ali–Frazier prizeFight, the "Thriller from Manila," seen at the two or three headends in the United States that had received 10-meter earth station antennas by then.

Figure 8.16 illustrates the major electronic components of the earth station used at cable headends. Signal is coupled from the "dish" antenna (see Chapter 7) to a feed located at the focal point of the antenna. The signal is amplified and, typically, down-converted to a lower frequency, by a low-noise block converter

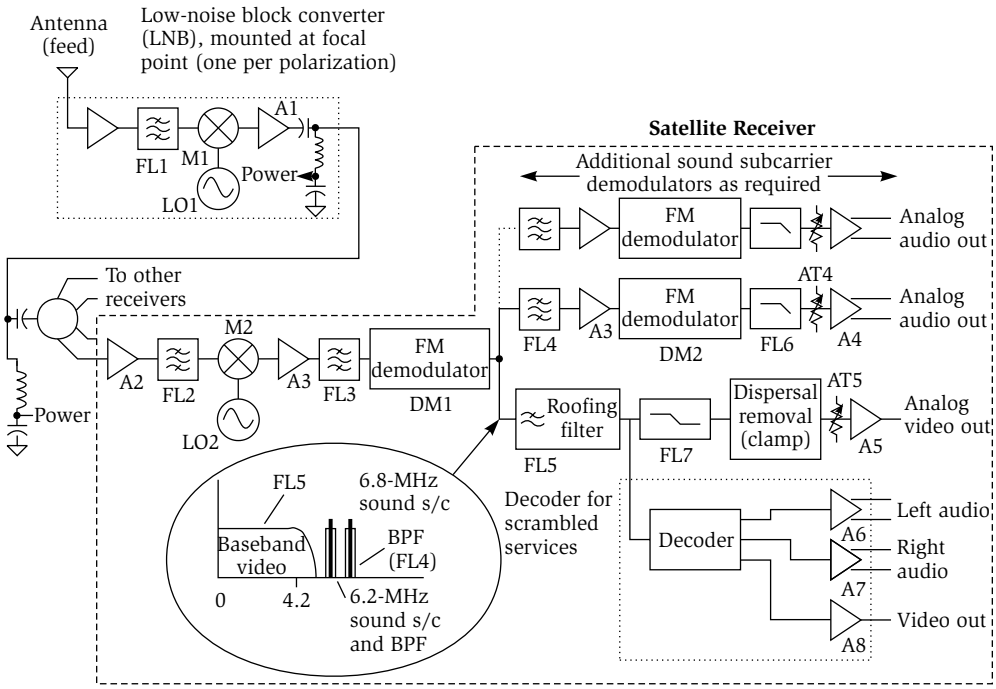


Figure 8.16 Analog earth station electronics, block diagram.

(LNB) located at the feedpoint. The LNB is located at the feedpoint to minimize signal loss. Communication satellite downlinks today operate on two bands, C band (3.7 to 4.2 GHz) and Ku band (11.7 to 12.2 GHz). Coaxial cable loss is significant at these frequencies though older earth stations used large, high-quality cable to route C band signals into the headend.

The most common band to which signals are converted is 950–1450 MHz (roughly L band). Some older equipment used 270–770 MHz. At these frequencies, the signal can be routed into the headend using fairly small coaxial cable. The cable may also carry power to the LNB as shown. Fiber optics may also be used to bring signals to the headend.

Once in the headend, the signal is split to serve a number of receivers: today, up to 12 receivers may be connected to a single LNB. A second LNB located at the same feedpoint can handle up to 12 additional channels coming from the same satellite in the same band, with different polarization.

At each receiver, the signal is again amplified by amplifier A2, filtered, and down-converted to another IF (dual conversion downconversion will be used if the receiver is frequency agile). The downconversion is done in mixer M2, using local oscillator LO2. After amplification and filtering, the signal is demodulated in FM demodulator DM1. As shown in the insert of Figure 8.16, the spectrum coming out of DM1 consists of baseband video (probably scrambled) and possibly one or more sound subcarriers. A number of subcarrier frequencies have been used for sound, but the two most common are 6.2 and 6.8 MHz. Today, these subcarriers are generally not used for primary program audio, that being carried digitally in the video. The subcarriers may be used for cue tones and supplementary audio. Cue tones are used to initiate local advertising inserts.

If an aural subcarrier is used, it is filtered in bandpass filter FL4, amplified, and detected in FM demodulator DM2. The audio is preemphasized using the standard 75-μs preemphasis, and so must be deemphasized in FL6. Attenuator AT4 allows the output level to be set, and differential output amplifier A4 supplies the output gain.

In the video path, a roofing filter FL5 may be used to remove power above the video spectrum. If not scrambled (rare today), the video passes from the roofing filter to deemphasis filter FL7. The video is emphasized to reduce the effects of noise on the signal, as is audio. A particular emphasis and deemphasis curve is used, as prescribed by the CCIR. After deemphasis, a dispersal removal clamp removes any dispersal signal that may be imposed on the video. Dispersal is used in C band satellite circuits because the same 3.7–4.2 GHz band used for satellite downlinks is also used for terrestrial microwave communications, primarily by the telephone industry. The fear is that if modulation is removed from the satellite signal, enough power could be concentrated at one frequency to interfere with terrestrial transmissions. A dispersal signal was required in the early days to prevent that from happening. The dispersal signal is a 30-Hz triangular waveform synchronized to the vertical blanking interval (VBI) and changing direction during the VBI. It is removed using a clamp or dc restorer,

somewhat as illustrated by C1 and D1 of Figure 8.4. From the deemphasis filter, the video level is set using attenuator AT5. After buffering in A5, nonscrambled video is output from the receiver.

Dispersal waveforms are not always used in modern practice. If a transponder is used full-time for video, then the downlink power will not be concentrated at one frequency. For transponders having occasional video service, dispersal waveforms are still used, but they are frequently used only to replace video when it is removed but the uplink remains on the air (possibly for test purposes).

Almost all signals delivered to cable headends today are scrambled. Programmers went to scrambling during the 1980s to prevent theft of service by privately owned earth stations that had not paid for service. The analog scrambling system used almost universally in the cable industry today is a sync replacement system (similar to sync suppression — see Chapter 21), with the horizontal blanking interval (HBI) replaced by encrypted audio. The idea is to make the video scrambling hard to break. Then the audio is "hard" encrypted, such that it could not be decoded without authorization. Descrambled audio and video outputs from the receiver are amplified in A6 through A8, and supplied to the outside, normally to modulators and stereo encoders as described earlier.

Signals to be descrambled are passed to the decoder, which accepts authorization and then descrambles the signals. The system in use has the capability to deliver two audio signals so that a stereo pair can be transmitted. If a second language program is transmitted, it uses one of the FM subcarriers described earlier.

8.7.1 Digital Satellite Receivers

Digital satellite receivers look similar to analog receivers at the block diagram level, with the exception that the demodulator is QPSK, OQPSK, or sometimes 8-PSK. All of these modulation formats are defined in Chapter 4. Typical receivers include an MPEG decoder, and they supply both analog video (from one channel in the received multiplex) with stereo sound and the entire digital transport stream in digital form. Audio outputs may include analog and sometimes a professional digital audio interface.

Figure 8.17 illustrates a digital earth station electronics block diagram; compare it with Figure 8.16 for the analog equivalent. Note that the antenna and LNB are identical, as is the front end of the receiver. In fact, the same antenna and LNB can be used to receive analog on some transponders and digital on others. The only difference is in the receiver.

As shown, the front end of the receiver is the same as in the analog receiver, but rather than an FM demodulator, a QPSK demodulator (most common today) is used. Its output supplies a descrambler that removes the signal scrambling done at the uplink to protect the programs for being pirated. Commonly, an analog decoder is supplied through a program selector, to recover an analog signal from one of the signals being transmitted. Its output is normal analog audio and video. An ASI driver formats the digital video signal for output, as described shortly.

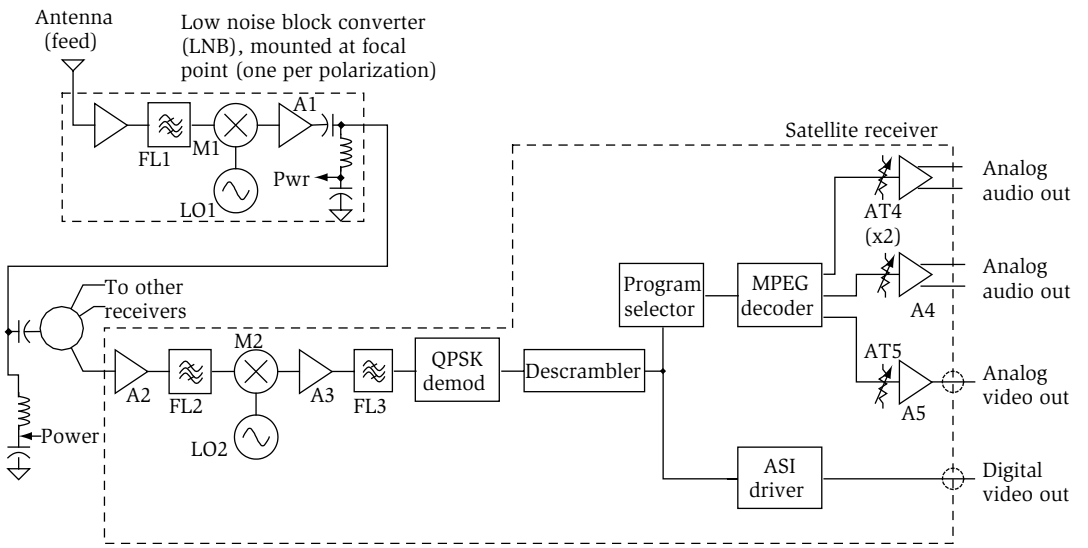


Figure 8.17 Digital earth station electronics, block diagram.

8.8 Digital Video Interfaces

Digital video interfaces connect one or (usually) more MPEG-2 transport streams (TS) from source equipment, such as a satellite receiver, an encoder, or a video-on-demand (VOD) server, to information sink equipment, most commonly QAM modulators. The MPEG-2 transport stream is described in Chapter 3. It includes video and audio as well as ancillary data for one or more programs. Note that multiple audio streams may be associated with one video stream. These audio streams can include surround sound and one or more languages. Since the audio, video, and ancillary data are transmitted in the same transport stream, there is only one connection for a program; there are no separate audio and video connections as you have in analog transmission. Also, it is most common to have more than one program in the same transport stream.

8.8.1 Asynchronous Serial Interface (ASI)

A common interface for digital video is the *asynchronous serial interface* (ASI), specified by the DVB Project Office. This is a serial interface, as the name implies, transmitted over a single 75-ohm coaxial cable. Besides coaxial cable, there exists a standard for an ASI using fiber-optic cable.

ASI is a one-way data transmission standard, with no acknowledgment of receipt of information and no built-in error correction. Being derived from the fiber channel standard, the signal is 8B/10B encoded, as explained in Chapter 19. For the coaxial cable version of the ASI, the transmitter supplies a baseband digital signal at an amplitude of 0.8 volts p–p into 75 ohms. The receiver must successfully detect a signal amplitude of 0.2 volts p–p. The return loss of the 75-ohm system is ideally 17 dB minimum.

The wire speed (actual data speed) of the data is 270 Mb/s with 8B/10B coding. Removing the 8B/10B coding results in a data speed of 80% of this, or 216 Mb/s. Included is a layer of ASI sync. One ASI link is able to contain enough data to fill several channels with 64- or 256-QAM data.

Figure 8.18 illustrates the logical ASI. The input at the transmitter (internal to a piece of equipment, so you normally won't see it) is a packet-synchronous MPEG-2 transport stream. As it enters the ASI transmitter, the signal is 8B/10B encoded. Chapter 19 explains 8B/10B encoding. Since the MPEG signal does not operate at the ASI data rate, there will be times when the ASI transmitter is ready to send more data and yet no data is ready to be sent. At such times, *filler code* words are inserted. This must be done to keep a constant bit rate on the ASI link. The signal is converted from parallel to serial for transmission, amplified and impedance matched, and then coupled to the connector. Transformer coupling is anticipated in the specification.

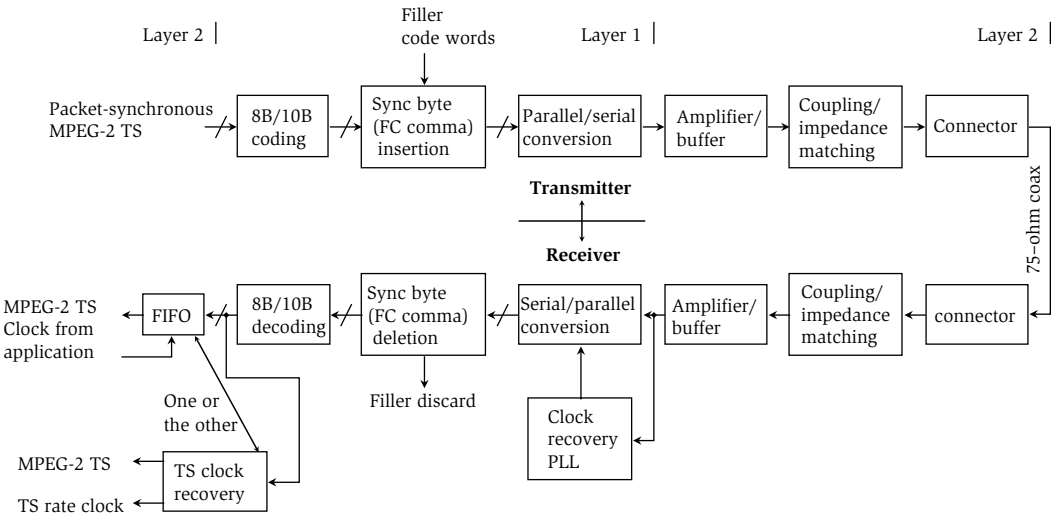


Figure 8.18 Digital video ASI.

After transmission, the signal is coupled into an amplifier through the input-coupling and impedance-matching mechanism. The bit clock can be recovered from the transmitted signal, and controls conversion of the signal from serial back to parallel for further processing. The ASI sync is detected, and the filler code words added in the transmitter are removed. 8B/10B coding is removed. The receiving circuit has a choice of how to recover the transmission stream data sync. Either the signal can be reclocked from the receiving application's clock or the transmission stream clock from the incoming signal may be recovered and used in the receiving application. In the former case, the signal is supplied to a first-in-first-out (FIFO) register, which is used for retiming a signal. It is clocked in from the transmission stream's clock and clocked out from the application's clock. In the latter case, the transmission stream clock is recovered and supplied to the application, along with the transmission stream itself.[6,7]

8.8.2 Ethernet for Video Transmission

Gigabit Ethernet is also popular as a method for transferring digital video in headends. The primary advantage is speed. ASI has a practical transfer capacity of about 160 Mb/s after all inefficiencies and overhead are removed. On the other hand, the practical transfer capacity of a dedicated gigabit Ethernet link can be as high as 900 Mb/s.[8] This means that when coupling signals between high-density video sources and modulators, such as in a video-on-demand system, fewer interfaces are required with gigabit Ethernet interfaces than with ASI. Recall that there does not have to be a 1:1 correspondence between headend interface cables and channels; one headend interface can serve several 6-MHz channels' worth of video, so long as the video is coming from one place (e.g., a VOD server) and going to one place (e.g., a cluster of QAM modulators).

When gigabit Ethernet is used for video transmission, it is common to embed the MPEG-2 transmission stream into RTP packets (explained in Chapter 6 — this same packetization is done for VoIP). The RTP packets allow for removal of jitter introduced in the transmission process; they also keep track of packet numbers to allow reordering of packets received out of order. (Out-of-order packets should not be a problem for local point-to-point transmission of video.) The RTP packets are embedded in UDP packets, which enable one-way or broadcast transmission of packets. In turn, the UDP packets are embedded in IP packets, as explained in Chapter 5, and the IP packets are embedded in Ethernet packets.

8.9 Signal Handling in Headends

This section includes information that should be useful to practitioners setting up and maintaining headends. Included are notes about video and audio signal handling. RF signal handling is covered in more detail in Chapter 9.

8.9.1 Video Loop-Through

Figure 8.4 included the illustration of video loop-through. This is a technique whereby you can supply video to several using devices such as modulators, video monitors, or switchers, without using amplifiers. It is normally recommended that no more than three devices be cascaded with a loop-through because there is adverse effect on the signal from each device. This section explains what loop-through is and the most common alternative to it.

Figure 8.19 illustrates video loop-through. Figure 8.19(a) illustrates video being looped through three modulators. The devices could also be a video switcher, a picture monitor, and a modulator. A common application in headends is to loop the video through a BTSC stereo encoder, a scrambler, and then terminate it at a modulator. Each modulator in the example has the ability to either terminate the video in a 75-ohm resistor or not. It is necessary to terminate video in the characteristic impedance of the coaxial cable carrying video so as to prevent reflections that would damage the frequency response of the picture.

If a signal is not terminated, it will be reflected toward the source. This reflected signal will combine with the direct signal, resulting in a frequency

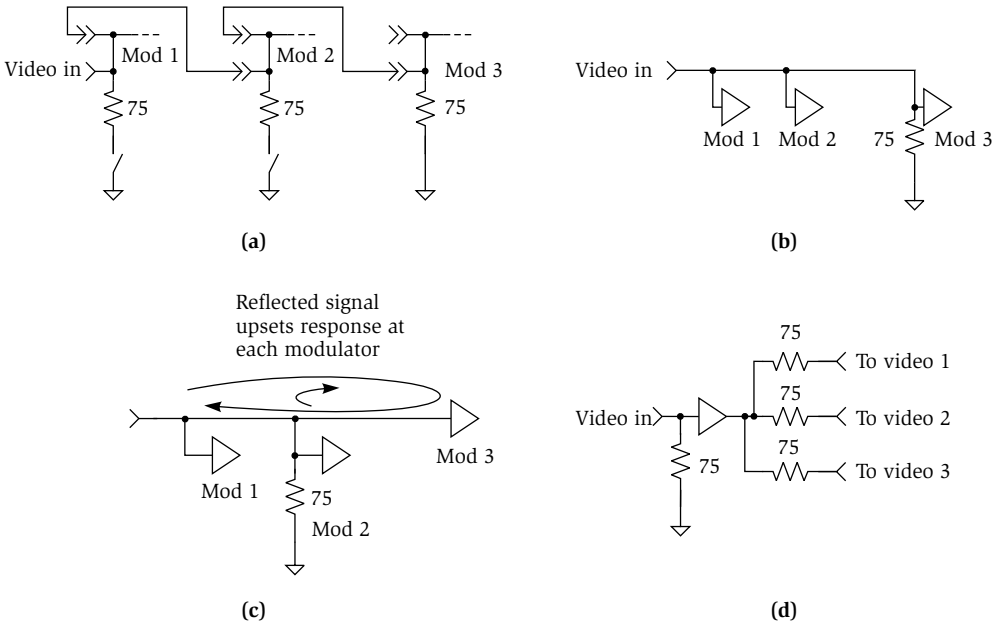


Figure 8.19 Video loop-through. (a) Looping through three devices. (b) Equivalent circuit for (a). (c) Incorrect application of loop-through. (d) Video distribution amplifier alternative.

response that peaks at frequencies where the two signals are in phase and nulls where the two cancel.

So long as the coax is terminated at the last device in a chain of devices, there is not a problem. This is illustrated in Figure 8.19(a), where three modulators or other devices are in a cascade, with the video being terminated by the resistor in the last device. The modulators shown all have the ability to terminate or not. This is usually selected by a switch on either the back or front of the modulator.

Figure 8.19(b) illustrates the equivalent circuit of the arrangement of Figure 8.19(a). High input impedance amplifiers are used to sample the signal from the coax transmission line without affecting significantly the impedance of that line. At the last modulator, a 75-ohm resistor terminates the line. As illustrated in Figure 8.19(c), if the termination is placed other than at the last device, reflections will result. Signal power will be reflected from the last device in line, modulator 3. When it arrives back at modulator 2, several bad things happen. First, depending on the frequency and the length of cable between the two modulators, signal components at some frequencies will add, increasing the amplitude, whereas others will cancel. The frequency response at modulator 2 is thus corrupted.

Furthermore, reflected energy arriving at modulator 2 will see a double termination: the 75-ohm terminating resistor in modulator 2 and the 75-ohm impedance of the transmission line leading back toward the source. This impedance mismatch creates a 6-dB return loss, so one-half the signal propagates toward the source (to the left), and one-half propagates again toward modulator 3, upsetting the response there. Thus, it is essential that a looped-through signal be terminated at the last, and only at the last, device in the signal chain.

Similarly, it is unacceptable to parallel more than one signal path because reflections will cause frequency response problems. This would happen if you branched off from device 1 and routed signals to devices 2 and 3 on different cables.

Each device the signal is looped through must use a very high impedance amplifier to bridge the video signal. As a practical matter, it is not possible to prevent some signal degradation as you loop video from one device to another. Because of this, the number of devices a signal is looped through should be limited. A rule of thumb is that three devices can be looped through without excessive deterioration in signal quality. Another limitation is that if a device the signal is looped through must be removed, then service will be interrupted to other devices in the loop-through chain. Finally, looping through may result in awkward headend wiring. An alternative, when video is to be distributed to many locations, is to use a video distribution amplifier (VDA).

Figure 8.19(d) illustrates a common form of VDA though most have more than three outputs. An amplifier provides a voltage gain of 2 to the video signal so that a 1-V p-p signal at the input becomes a 2-V p-p signal at the output. A number of outputs are provided, each with a series 75-ohm resistor to establish

the source impedance of the path. This resistor and the 75-ohm terminating resistor in the following device form a voltage divider, reducing the voltage back to the video standard interface of 1-V p-p.

8.9.2 Audio Wiring

Many of the same factors that apply to video also apply to audio, except that audio is usually less critical in terms of signal wiring than is video because of the lower bandwidth. It is possible, and many times recommended, that audio distribution amplifiers (ADAs) be used for much the same reasons that VDAs are used. However, it is sometimes acceptable to wire a number of audio-using devices in parallel without regard for reflections since typical headend wiring distances are insignificant compared with the wavelength of audio. It is important, though, to provide one and only one 600-ohm termination in an audio chain.

Professional audio sources are usually specified to drive 600-ohm loads. Some sources may actually exhibit an output resistance of 600 ohms though it is not at all uncommon for the source resistance to be less. Since audio lines are usually of insignificant length compared with a wavelength, this lower driving resistance is not a problem unless it is desired to drive a signal several miles. In this case, more care should be taken to match resistance on both ends.

A major headache in audio wiring is hum elimination. Most hum is due to ground loops and other sources of potential difference. Ground loop mitigation is the most magic of all magic arts: no matter how many words are written about it, the problems encountered in the field will not be covered. However, some pieces of advice are sufficiently general to be worth mentioning.

Figure 8.20 illustrates options in audio interconnection between an audio source and load. Figure 8.20(a) represents the interconnection method strongly recommended and practiced almost exclusively in professional audio applications. Balanced transmission is used, in which the source consists of two signals having equal resistance to ground (not shown) and exactly opposite phase. One way to produce such signals is to use two amplifiers, one for the + phase and another inverting that signal to generate the − phase, as shown. Another way is to use an audio transformer.

The signal is connected to the load on the right of Figure 8.20(a), using a balanced, shielded cable. Several cables made for this purpose are available. The two wires of the cable are terminated in a differential amplifier, whose output is an amplified version of the difference between the two input signals on the + and − inputs. By using balanced transmission, any hum pickup will be the same on the two wires of the balanced pair, so the hum will be subtracted out at the receiving differential amplifier (or transformer).

Notice that the shield is connected to ground at only one end. It is not correct to ground the shield at both ends since this can create a ground loop.[9] Ground loops exist any time there is more than one path for current between pieces of equipment. As a practical matter, this is hard to prevent: racks must be bonded

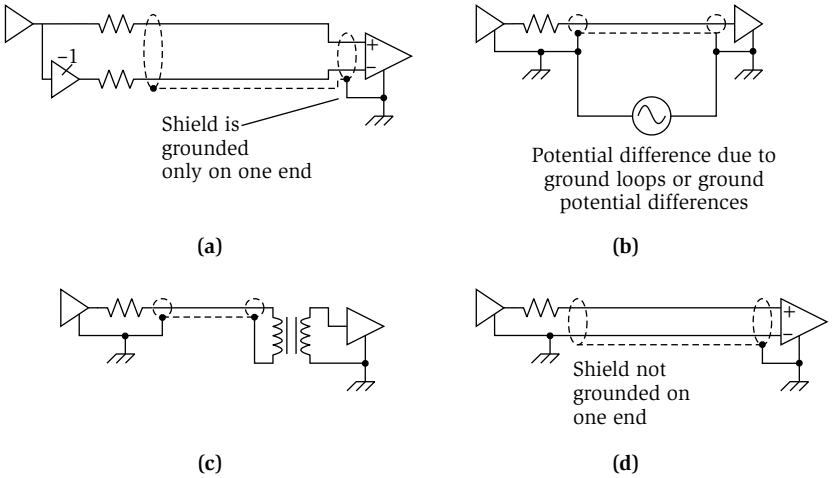


Figure 8.20 Options in audio interconnection. (a) Balanced audio. (b) Unbalanced audio with hum source. (c) Unbalanced source and load with transformer. (d) Unbalanced source and balanced load.

together and to ground, providing a path. Individual pieces of equipment must have third-wire safety grounds, providing a second path. After that, consider all the interconnecting wiring, much of which involves grounds. A loop can form the secondary of a giant transformer, the primary of which is all the ac wiring in the headend. The result is current induced in the ground paths, which sets up potential differences between equipment. Besides the transformer model, it is likely that many headends will experience potential differences between racks because of the various routes supply current can take to get back to the neutral of the primary power supply.

Figure 8.20(b) illustrates the unbalanced signal connection method used in most consumer equipment but not recommended for professional use. A single-ended output amplifier sends the signal to the receiving equipment, which receives it using a single-ended input. The shield of the cable serves as the return path. Also shown is a prime reason to not use this method in headends. If an ac voltage exists between the two ends, it will result in current on the shield. Because the shield has some resistance, a voltage drop will exist between the two ends. This voltage will be seen by the receiving amplifier, resulting in hum in the audio.

If the equipment at hand can accommodate only unbalanced audio, and if hum is a problem, then the solution of Figure 8.20(c) will have to be used. A high-quality audio transformer having a 1:1 turns ratio can be used at the receive side to isolate the ground on the two sides. The input side is connected to

the center conductor and shield, but not to ground. The output side is connected to the amplifier input and ground. Transformers *must* be operated at their specified impedance. Small, low-cost transformers will distort the audio, particularly at lower frequencies, though high-quality professional transformers are available that do a good job.

Figure 8.20(d) illustrates an option where the sending equipment is single ended, but the receiving equipment can accommodate a balanced input. Balanced audio cable is used, with the sending ground connected to one of the wires of the pair. Again, note that the shield is connected at one end only.

In the end, though, we can say only that ground loops are so different from each other, and so difficult to resolve, that whatever works is the right thing to do.

8.9.3 Controlling Hum in Video

Hum in video systems using NTSC and in 60-Hz power grids will show up as horizontal bars moving slowly up through the picture. The upward movement is due to the slightly slower than 60-Hz field rate of NTSC. Just as ground loops can cause great anguish with technologists attempting to interconnect audio circuits, it can be a problem with video. Because of the frequencies involved, balanced (differential) coupling of video is not as predominant as it is in audio applications though it is possible to interconnect video signals using balanced transmission. It is far more common to use single-ended equipment to interconnect video. This makes resolution of ground loop problems even more difficult for video than for audio.

Figure 8.21 illustrates two methods that can be used to mitigate hum in video circuits. Shown are video source and terminating equipment. Both the signal and chassis grounds are shown, emphasizing that they are usually one and the same in video equipment. They may be kept separate internally to the equipment, but for the most part, they are the same outside the equipment.

Figure 8.21(a) illustrates use of a hum bucking coil. These may be purchased commercially and consist of a number of turns of 75-ohm coax through a high-permeability ferrite core. Typically, BNC connectors are provided on the case of the hum bucking coil. The connectors must be insulated from each other. To a current on either the center conductor or shield, the arrangement presents a high inductive reactance, which opposes the current. However, to a current traveling in one direction on the center conductor and the opposite direction on the shield (correct differential mode signal transmission), the inductive reactance is not present: the inductance operates only on the net current in the coax, and in this case, the net current is zero.

So long as the inductive reactance at 60 Hz is high enough, it is not possible for significant ground loop current to develop through the coax. Thus, the hum bucking coil can offer significant reduction in hum, but only if it is not shorted between the input and output shields. The short would afford ground current a way to avoid the inductive reactance of the coil.

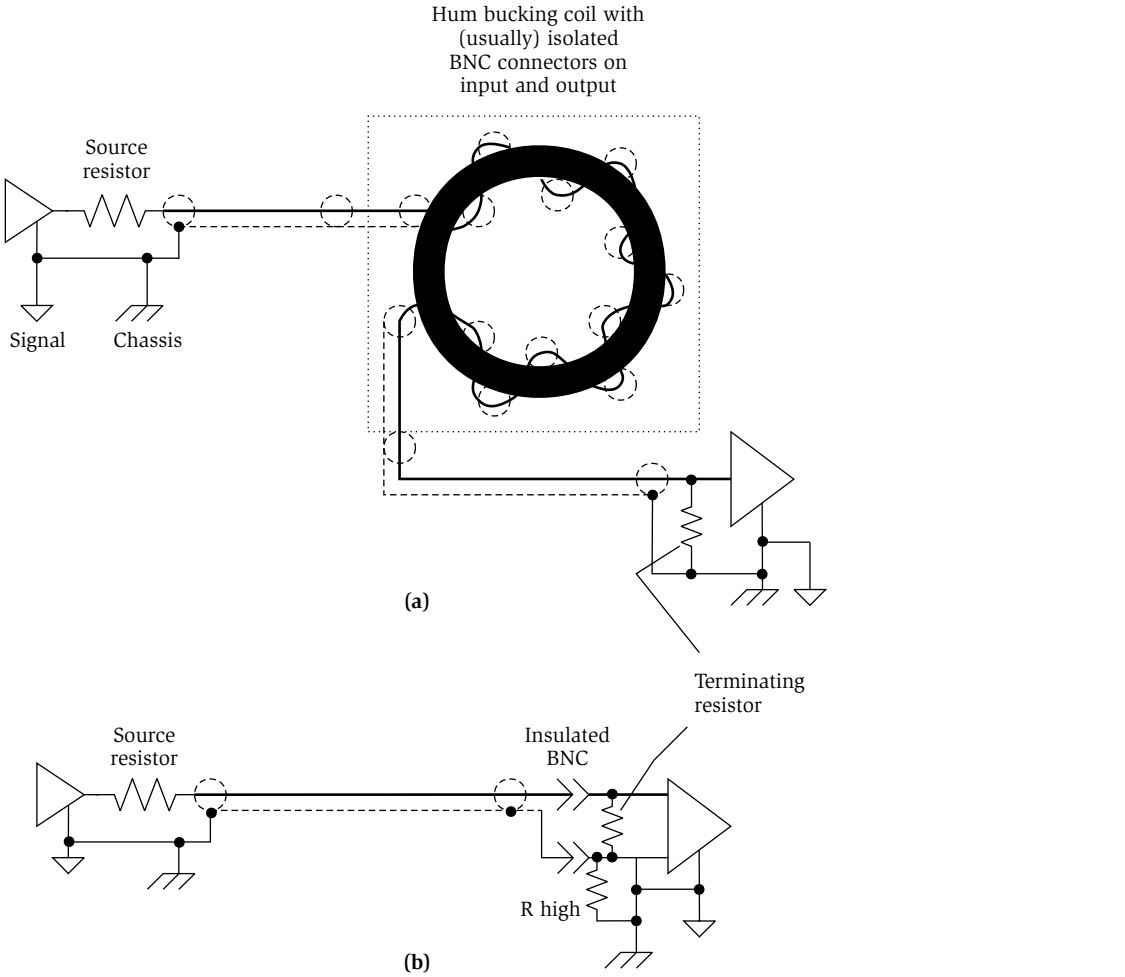


Figure 8.21 Methods used to mitigate hum in video circuits. (a) Hum bucking coil. (b) Floating input connector.

Some video equipment does offer a balanced input driven from an unbalanced source, as illustrated in Figure 8.21(b). This is analogous to the situation of Figure 8.20(d). Usually, a special BNC connector is used, which has its shield insulated from the chassis. A differential amplifier recovers the voltage difference between the center conductor and the shield. The resistor to ground is high and is used to prevent a static charge buildup on the shield if the other end is not terminated. A floating BNC connector as shown is used on some professional

equipment, but the cost prevents its use in lower-cost lines of videoprocessing equipment.

8.9.4 Handling Digital Video Signals in the Headend

Asynchronous serial interfaces should be routed point-to-point and not looped through from one piece of equipment to another, even if the same signals are bound for more than one place. Use a digital distribution amplifier to route the signals to more than one place.

ASI signals are handled on the same type of 75-ohm coaxial cable used for RF and analog video, so the same handling precautions apply. The digital signal will in general be more immune to interference pickup than is the analog signal. However, signal radiation from a digital interface to an analog interface is a possibility. The same measures taken to minimize interference in analog signal transmission will work here. Use the best-quality coaxial cable possible. Quad shielded cable is recommended for headend use. Pay particular attention to the shield connections, because poor shield connections can encourage signal radiation. If digital pickup in analog cabling is a problem, also try physically separating the cables. If this fails, try ferrite beads on both the digital and the analog cables. Make sure the ferrite beads are specified for the frequency range over which the interference is a problem. A ferrite optimized for use at power frequencies will perform very poorly at several hundred megahertz.

A lot of signal transmission is on Ethernet cables today. These cables consist of a number of twisted pairs handling balanced signal transmission. As shown in the telephony chapter, balanced transmission is a good way to minimize pickup or radiation of signals, but it is not perfect. If you are having problems with radiation from Ethernet cables, you can also try appropriate ferrite beads on the Ethernet cables. You can also try shielded Ethernet cables (STP). Many times, though, the best strategy is to separate the Ethernet cables from victim cables.

8.10 Ad Insertion

Insertion of advertising in the programs transmitted by various programmers is a significant source of income for many cable systems. Many basic cable networks charge cable systems for the privilege of carrying the network. In turn, the networks make available "avails," or time each hour that is available for sale of advertising by the local cable systems. Though practices vary, it is not uncommon that two blocks of time, each 2 minutes long, are made available at predetermined times each hour for insertion of advertising sold by the local cable system. Control of insertion is done by the network, which transmits cue tones to initiate the play of local commercials and cue tones to transfer back to the network.

In the past, ad insertion was done using analog tape players controlled by a computer. Today, file servers have replaced the tape players in new systems. Figure 8.22 illustrates a file server used for this purpose. Differences will exist from one system to another. The file server itself is usually a computer workstation optimized for transfer of large digital files from its internal hard disk drives to the outside world. It interfaces with a traffic system, which provides the schedule for playing commercials. Through a second interface, commercials are loaded onto the hard drives. Typically, the commercials are compressed using MPEG-2 encoding.

One file server is typically capable of supplying data fast enough to service a number of channels simultaneously. As shown in Figure 8.22, the file server "plays" the file out through a system interface, to an individual analog channel interface, which is shown in more detail in Figure 8.23. The channel interface receives analog video and audio from the IRD or other source, along with cue tones. Usually, the cue tones are transmitted on analog sound subcarriers since the audio is transmitted digitally with the video. FM demodulators, such as DM2

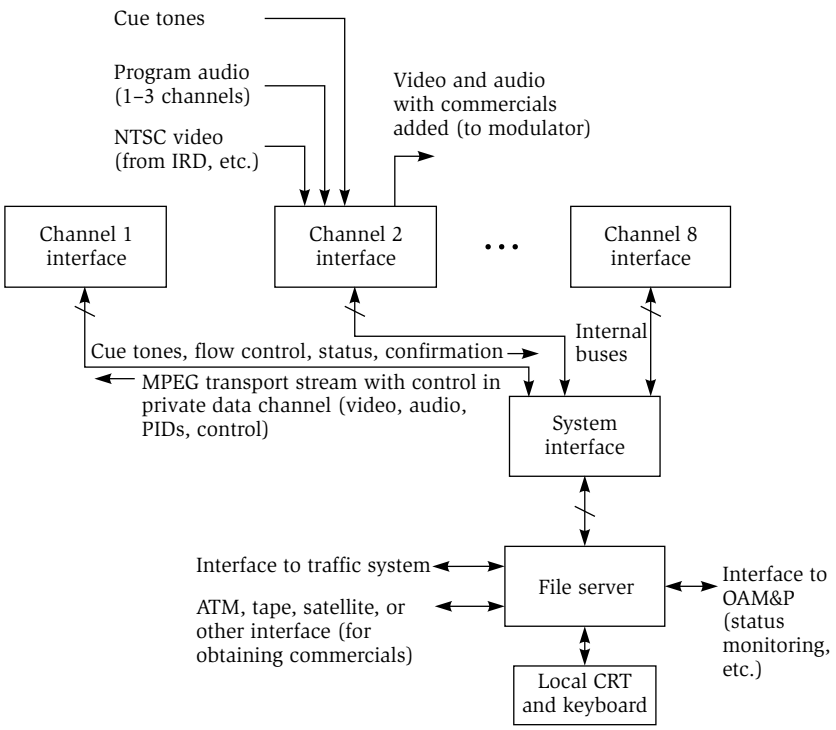


Figure 8.22 Video file server system used for ad insertion.

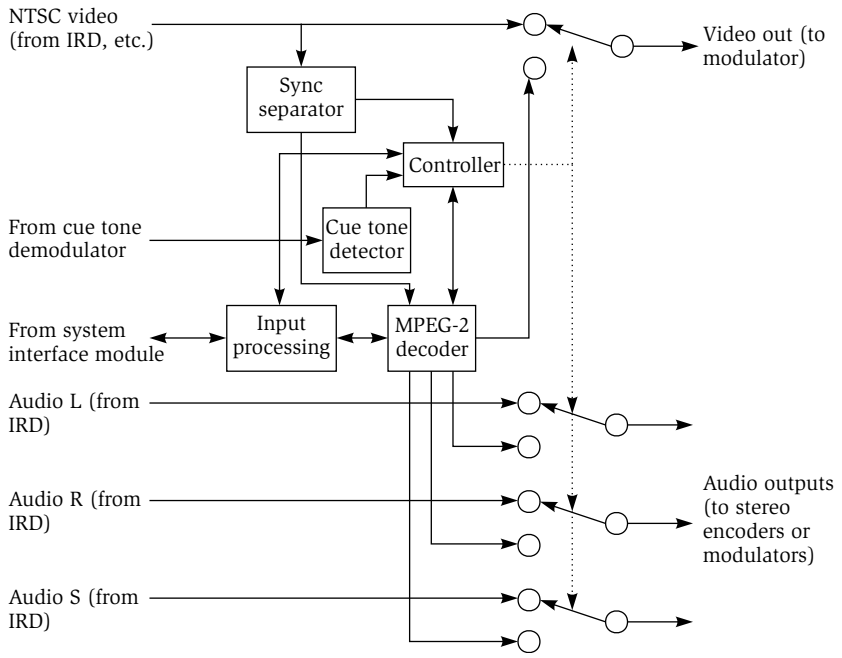


Figure 8.23 Individual channel interface in ad insertion system.

of Figure 8.16, recover the cue tones. The cue tones are passed to the file server, where they initiate playing of commercials.

Figure 8.23 illustrates an individual channel interface. NTSC video is supplied to a switch, which allows replacement of the incoming video with the commercial. Sync is extracted from the video to allow the inserted video to be synchronized with the incoming video, a requirement for smooth switching between sources. An MPEG decoder receives video from the file server and converts it to NTSC video. The cue tones are supplied to a detector, which passes them on, ultimately to the file server. One to three (if SAP is included) channels of audio are switched with the video.

The video and audio from the ad insertion system are supplied to a modulator and, usually, stereo encoder, for transmission to subscribers. If the inserted video is synchronized properly with the incoming video, and the actual switchover between incoming and advertising video is done in the vertical blanking interval, the subscriber does not see a change that indicates the video is being originated locally. Cue tones cannot be heard because they are transmitted on a channel other than the one used for program audio.

Ad insertion in digital programs is covered next.

8.11 Video On Demand and Digital Ad Insertion

Video on demand (VOD) is a popular service. VOD allows a subscriber to order his or her own custom video program from the headend. It is the cable equivalent of the video rental store, without late fees and rewind fees! Programs are stored on a server and listed in the electronic program guide (EPG) on the subscriber's STT. When a subscriber orders a VOD program, the request is relayed to the headend on the return data channel from the STT. The VOD server must determine a free time slot in a free channel on that subscriber's node. The STT is told what RF channel to tune to and what PID to select on that channel. The server then "plays out" the program, multiplexing it with several other signals bound for other subscribers on that node.

VOD is a personal service sent individually to one subscriber. However, in order to reach that one subscriber, the program must be sent to all subscribers on that fiber node. Only the one subscriber for whom the program is bound has the correct descrambling key to allow recovery of the program. Usually the subscriber has the ability to perform "VCR-like" functions, such as pause, rewind, and fast-forward. The subscriber is allotted so many hours to watch the program, an amount of time somewhat longer than the run time of the program. This will allow for meaningful pausing and rewinding.

Sometimes two or more levels of file server hierarchy are employed, with a file server in a hub close to a group of subscribers that has the more popular titles on it. A file server in the master headend or some other location keeps less popular programming and supplies it when requested. A significant issue in VOD is the need to accurately keep track of what programming is available and where it is located. The program owners may have strong feelings about controlling where the programming resides and how it is distributed from there.

A VOD file server can be constructed architecturally the same as an ad insertion file server, except more storage capacity is required. To improve reliability, VOD servers may employ *redundant array of independent (or inexpensive) disks* (RAID) technology, a category of disk drives that employ two or more drives in combination for fault tolerance and performance. In RAID systems, multiple disk drives are used to store different portions of the program data such that the failure of one disk drive does not cause the loss of any program material. There are several levels of RAID drives, of which levels 0, 3, and 5 are the most popular.

Level 0 provides *data striping* (spreading out blocks of each file across multiple disks) but no redundancy. This improves performance but does not deliver fault tolerance. Level 3 is the same as Level 0 but also reserves one dedicated disk for error-correction data. It provides good performance and some level of fault tolerance. Level 5 provides data striping at the byte level and also stripe error-correction information. This results in excellent performance and good fault tolerance.[10] In other words, RAID spreads out the bits of each byte in the file over several disk drives and puts the data back in the right sequence upon play out.

Error correction is added to allow the file to be recovered if any one disk drive fails. Typically, the disk drives are hot-swappable, meaning you can change a failed drive out without stopping operation of the file server. The server can operate with one disk down by doing on-the-fly error correction. As soon as the faulty disk drive is replaced, it can rebuild the file by error-correcting the surviving data.

A related service is a network *personal video recorder* (PVR). The idea is to use a video file server to capture programs as they are broadcast and then to allow subscribers to view them later as a VOD program.

8.11.1 Digital Video Rate Management

Cable systems have long inserted advertising on analog channels. But with the advent of digital television, ad insertion had to be extended to this medium. In principle, the block diagram of Figure 8.23 applies to insertion in digital video streams, but there are some new issues to consider. The first issue in digital ad insertion is that you have to exit the programming stream in favor of the advertising stream. You do this once you have transmitted enough information to recover all frames of the program picture that are to be displayed prior to starting the advertisement. Then you must start the advertisement with an MPEG I-frame, since it is the only frame that will contain enough information to allow the first frame of the advertisement to be displayed. It is not enough to do this just for video; you must make the audio match what the video is doing, and audio is compressed independent of the video. Typically, the audio will be muted for a short time during the switch-over. Finally, since the advertisement is being inserted into a multiplex of channels that have usually been statistically multiplexed, you have to ensure that introduction of the advertisement does not overrun the capacity of the RF channel.

The first job, digital program insertion, is covered in Section 3.5.4. Figure 3.26 shows the steps that must take place to leave the normal program stream and go to the ad insertion stream. The second issue, managing the data rate to stay within the available bandwidth, is covered here. This discussion applies any time you modify the content of a digital video stream in a VBR environment (see next subsection). This happens when you insert a local commercial and when you groom an incoming video feed to select some programs and combine with other programs on a different incoming stream.

Constant and Variable Bit Rate

Video programs may be transmitted using either a *constant bit rate* (CBR) or a *variable bit rate* (VBR). The terms are self-explanatory. If a program is sent using CBR, the data rate is set to whatever is necessary to provide a good picture at all times. If fewer bits are required, then filler bits are added ("stuffed") to round out the set bit rate. If by chance, however, a picture would benefit from a higher bit rate for a time, the picture can't have a higher bit rate; the encoder must make do with the bit rate allocated.

VBR transmission is more efficient because it transmits only whatever data rate is necessary to recover a good picture. VBR transmission frequently originates where video is encoded, using a *statistical multiplexer*, or stat mux. A stat mux monitors all the data coming from a number of encoders, which data is being multiplexed into one transmission stream. The transmission stream, of course, is at a constant bit rate, required to enable the receiver to lock to the transmission. But the transmission stream can be composed of a number of VBR streams managed by the stat mux. The stat mux buffers each encoder's output, multiplexing all outputs into the output transmission stream. If the buffer starts to empty, meaning that not enough bits are going to be available to make up the output transmission stream, then one of three things can happen: (1) The stat mux can call for more data from one or more of the encoders, even if the additional data is not needed for quality video; (2) the data stream is stuffed with filler bits; (3) opportunistic data is transmitted, that is, data held in queue until there is room for it.

On the other hand, there will be times when the stat mux buffer starts to get too full, indicating that there is not going to be room in the transmission stream for all of the bits. In that case, the stat mux must tell one or more encoders to back off and encode to a lower bit rate, even if the target video quality is compromised. Hopefully, the data peak is over in a fraction of a second, and viewers remain unaware that an encoder had to compromise its target quality momentarily.

Most VOD systems store material in CBR format, though there could be savings of maybe 30% on storage space by using VBR.[11] The difficulty with VBR is that when you multiplex several VOD streams together, you have a problem managing the data rate. This job is simplified if you work with CBR.

Many program multiplexes are encoded using VBR, however, in order to gain efficiency in transmission. The problem comes in when you pull some program streams out of a VBR transmission stream and combine them with other VBR streams, from an ad insertion system, from a broadcaster, or from somewhere else where you have no control over the data rate to which the material was encoded.

Figure 8.24 illustrates a VBR transmission stream with four programs. In order to provide better quality for programs that we want to look better, we have set a peak rate for program 2 that is three times that of programs 1 and 4, and we have assigned a peak rate for program 3 that is twice that of programs 1 and 4. You can see that in two instances in the time monitored, the total rate for the four programs exceeds the capacity of the channel. If the programs were being put together by a stat mux, it would have one or more encoders back off during the peaks. You can also see several valleys in the total rate, in which opportunistic data may be inserted if available.

The problem comes in where you have assembled this transmission stream from several independently encoded programs and you have no control over the incoming data rate of each. Suppose, for example, that program 4 is a commercial you have inserted into a transmission stream. You have no control of the three program streams you are passing through, and the commercial substituted for program 4 is preencoded, so you have no capability to influence its data rate

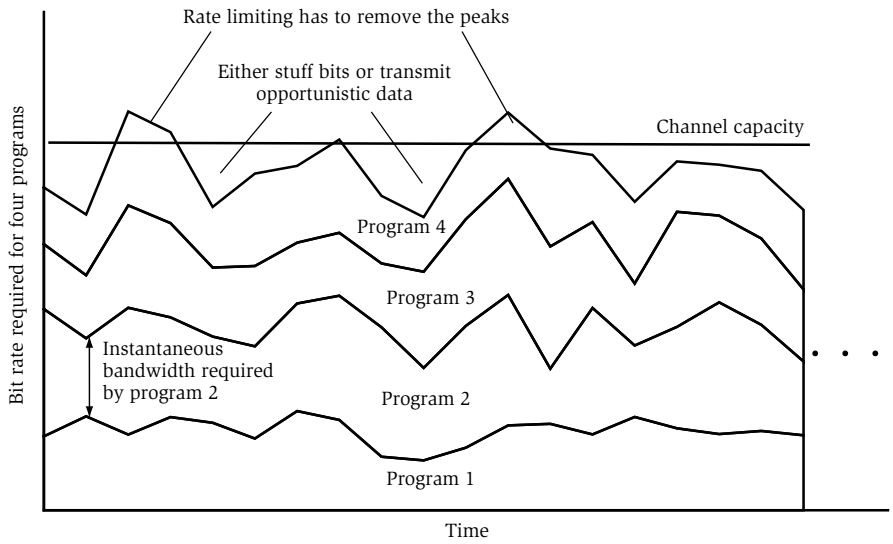


Figure 8.24 VBR transmission stream with four programs.

either. Specialized equipment called a *rate changer*, or a *groomer*, has been developed for this situation.

It is possible to decode each program back to the non-MPEG-2 encoded video, and to reencode using a stat mux system. There are a couple of difficulties with doing so. The first is that it is rather expensive, and it must be done with each program in the multiplex. The other problem is that, as pointed out in Chapter 3, MPEG-2 encoding is a lossy process. If you cascade multiple encoding, you are going to damage the picture. It is possible to find that the second encoding process needs a higher bit rate to handle artifacts introduced in the first encoding process.

In order to avoid these problems, a rate changer analyzes the picture and makes a judgment on what information can be dropped from the program stream without doing noticeable damage to the picture. It may be that the rate changer will decide that it doesn't need to encode the higher-order DCT coefficients with as many bits as was done originally, or it may decide that it can drop some of the higher-order coefficients altogether, for example. Figure 3.9 and the associated text shows what is going on here.

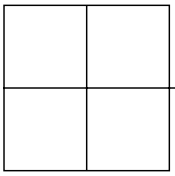
8.12 Summary

This chapter has provided information about the common types of equipment found today in headends. This equipment includes processors, modulators,

demodulators, stereo encoders, earth station receivers, and advertising insertion equipment. We have demonstrated typical circuit sets and features, recognizing that a number of ways to do each job exist, and that not every piece of hardware will have every feature shown. The next chapter will cover what happens to signals after they exit the individual pieces of equipment and will also discuss other techniques related to headend use.

Endnotes

1. www.wegener.com. See data sheets for DTV 700 Series.
2. Donald G. Fink, *Electronics Engineers' Handbook*. New York: McGraw-Hill, 1975, pp. 2–26.
3. Christopher Bowick, *The Importance of Setting and Maintaining Correct Signal and Modulation Levels in a CATV System Carrying BTSC Stereo Signals*. This has appeared in several places, but a useful source is the *BCT/E Certification Program Reference Bibliography Reprint Manual*, Society of Cable Telecommunications Engineers, 140 Philips Rd., Exton, PA 19341–1318. Telephone 800–542–5040. Many other useful reprints are also found in this volume.
4. IBA, BREMA, BBC, *NICAM 728: Specification for Two Additional Digital Sound Channels with System I Television*, 1988. The current status of this document cannot be determined. However, the Instituto de Communicagoes de Portugal has published substantially the same information, which can be found at http://www.icp.pt/legisuk/p316–93uk.htm.
5. Televerket Radio Laboratory (Sweden), *Digital Multisound in Television — Feasibility of Filtered 728-Kbit/s NICAM with 5.85-MHz Carrier Frequency in CCIR System B. Adjacent Channel Operation in Cable Networks*. Paper is dated January 16, 1987, and was privately furnished to the author. Publication information is unknown.
6. DVB Project Office, *Interfaces for CATV/SMATV Headends and Similar Professional Equipment*. DVB Document A010, rev. 1, 28 May 1997.
7. ETSI, *Digital Video Broadcasting (DVB); Professional Interfaces: Guidelines for the implementation and usage of the DVB Asynchronous Serial Interface (ASI)*. ETSI TR 101 891 b1.1.1 (2001–02).
8. Per private correspondence with Bob Gaydos, Concurrent Computer Corporation.
9. Linc Reed-Nickerson, Audio Levels Scream for Attention. *CED*, October 1997, p. 54ff.
10. www.webopedia.com
11. Per private correspondence with Bob Gaydos, Concurrent Computer Corporation.



Chapter 9
Headend Operation

9.1 Introduction

Previous chapters in the headend section have covered the input devices to headends and the electronics that compose the bulk of the headend equipment. This chapter covers a number of issues related to headend operation. First, the accepted band plan for cable television is presented. This defines the accepted channel number versus frequency plan adopted by both the cable television and consumer electronics industries, and mandated by the FCC.

Management of optical and RF connections and signals is important in today's headends, so these topics are covered in some detail. This topic includes steering signals to and from nodes and organizing wiring such that performance is optimized. Finally, topics related to the assessment of the quality of signals in headends are covered.

9.2 Cable Television Band Plan

For reasons shown later, the frequencies used for transmission of television signals are not contiguous, as they are in AM or FM radio broadcasting. Therefore, television stations are identified, not by frequency as is the case with radio, but by channel number. The channel numbers are easy for viewers to remember and avoid the confusion of noncontiguous frequency assignment.

9.2.1 Concept of Image Response and Historical Background

The VHF television assignments are chosen to allow certain simplifications in tuner design. Recall from Chapter 2 that the NTSC television signal modulated onto an RF carrier has a picture carrier, a color carrier located 3.58 MHz above the picture carrier, and a sound carrier located 4.5 MHz above. The spectrum from 0.75 MHz below the picture carrier to 4.2 MHz above the picture carrier (at RF) is occupied with visual information. Engineers often talk of the channel frequency as the frequency of the picture carrier, understanding that the other carriers and the modulation sidebands are also present.

Inside a demodulator or television receiver, the incoming signal is quickly converted to an intermediate frequency (IF). This is a standard frequency at which a tuned signal on any channel will be amplified, filtered, and detected. The spectrum of the signal at IF is reversed, or inverted, from that at RF, due to use of a high-side local oscillator (LO) in the TV receiver. Using a high-side rather than low-side LO mitigates certain problems that would create major performance limitations. Local oscillators are used to convert an incoming frequency to another frequency and sometimes are used to convert from one internal frequency to another. When you tune a TV or radio, you are changing the frequency of the first local oscillator such that the difference between its frequency and the desired incoming frequency equals the IF frequency.

Figure 9.1 illustrates the frequency conversion process, starting with the RF frequencies transmitted over the air or on cable and ending in the IF band normally used today. The top portion is taken from Figure 8.10 and shows the conversion of the received signal to IF. The lower portion illustrates what happens when a TV is tuned to channel 2. The local oscillator operates at 101 MHz for channel 2. The mixer takes the difference between the LO frequency (101 MHz) and the RF picture carrier (55.25 MHz). This difference is 45.75 MHz, the standard picture carrier IF used in NTSC television reception. Also shown are the conversions for the sound and color carriers. Notice that the spectrum of the channel at IF is inverted from that at RF, due to subtraction of the RF frequencies from the local oscillator frequency.

If bandpass filter BPF2 is not present, and if there exist incoming signals at cable channels 17 and 18, then an "image" is produced in the IF. The desired IF signal is produced by subtracting the incoming frequency from the LO frequency, but the image is produced by subtracting the LO frequency from the incoming image frequency. This is shown in the IF spectrum, where the channel 18 picture carrier is converted to 44.25 MHz by the same local oscillator that converts the desired signal. The presence of these "spurious" signals in the IF passband will produce interference beats. No amount of filtering at IF (represented by BPF3) can remove the interference. The only way to remove it is to never allow the image-producing signal to reach the mixer. It can be removed by filter BPF2 before frequency conversion. This filter is tuned to the desired channel.

The image frequency is given by $f_{\text{image}} = f_{RF} + 2f_{IF}$. It is desirable that the image frequency be removed as far as possible from the desired RF frequency to simplify BPF2. The filter must allow adequate rejection of the image frequency so that the image does not adversely affect the received signal. If the IF frequency is raised, then the transition region of BPF2 can get larger, resulting in a simpler filter. The transition region is the frequency region between where the filter must pass signal (the passband) and the frequency at which it must provide a specified rejection (the stopband).

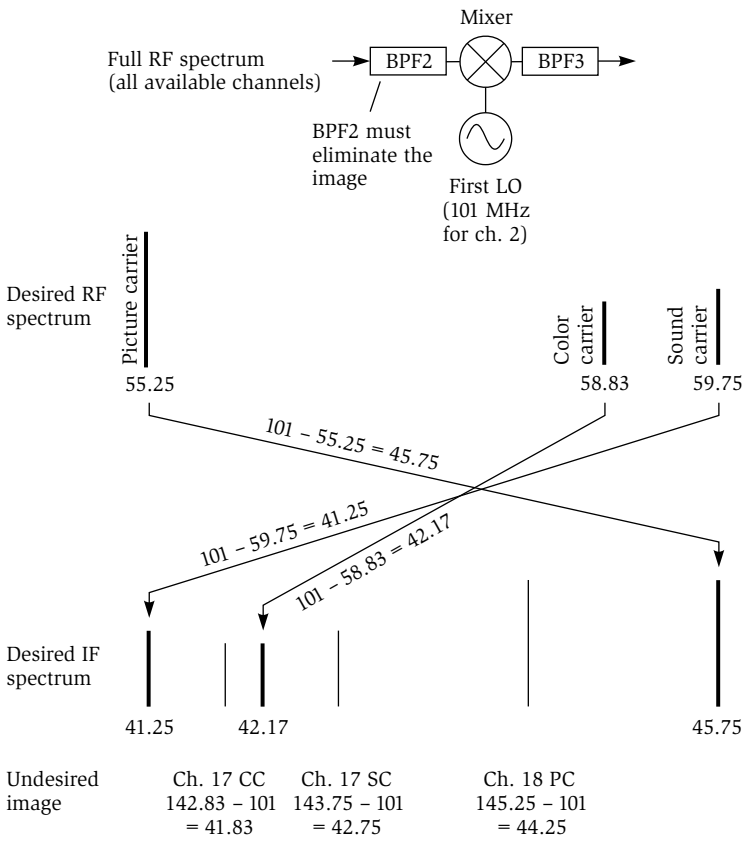


Figure 9.1 Operation of high-side local oscillator at channel 2.

VHF off-air allocations don't include the image frequency for any channel. Thus, a TV receiver designed for off-air reception doesn't have to worry about a television signal being present on the image frequency. Furthermore, local oscillator frequencies don't fall within a channel. If a set radiates power at the local oscillator frequency, there is no danger of causing interference to TV reception. The UHF band does not include these desirable characteristics since channels there are contiguous over a frequency range of 336 MHz. However, the channel allocation plan the FCC uses includes "taboo" channels. When one channel is assigned in a city, several other channels become unavailable owing to image, local oscillator, and adjacent channel considerations. Assignment of channels for digital television broadcasting violates the taboo channels for the first time in the history of North American television broadcasting.

9.2.2 Introduction to the Modern Cable Television Band Plan

The need to use spectrum efficiently has driven the cable industry to include channels that violate the off-air plan and taboos. That this has worked is due in part to the fact that signal levels in cable are controlled much better than they can be controlled over the air. When all else fails, set top terminals are used to combat image problems. Set-top terminals use a dual-conversion frequency plan with a first IF above the tuning range, so image rejection is not a problem.

Early on, the cable industry began using the so-called midband frequencies between the FM band and the VHF high band (channels 7–13). Initially, these were designated using letters, beginning with channel A (120–126 MHz). When need and technology opened up higher frequencies, the spectrum above channel 13 (210–216 MHz) was called the superband, and channel designations continued to use sequential letters. About 1980, set top terminals began using all-electronic tuning in order to permit the use of economical remote control. The low-cost numerical displays were not capable of displaying letters, so set top manufacturers began using numerical channel designations, beginning with channel 14 for what had been known as channel A.

Each manufacturer used his or her own allocation plan though to an extent a consensus developed: what had been channel A became channel 14, and sequential numbers were used from that point on. The problems developed when they got to channel 36, 298.25 MHz. This was the top frequency being used when mid- and superband numerical designations became popular. In order to get a little more capacity, cable systems were sometimes putting TV signals in the FM band, 88–108 MHz, and in the 108–120 MHz region previously avoided because of concern for potential aeronautical interference. In addition, HRC and IRC plans were becoming popular. As we will show, use of these plans opened up another 6-MHz block between channels 4 and 5. When set-top manufacturers added all these channels to set-top terminals, they took different directions in channel number assignment.

Shortly after, TV manufacturers began using tuners that tuned the cable channels. With such tuners and 75-ohm inputs, the sets were called "cable ready" and were sold as not needing set-top terminals.* Lacking a standard channel assignment plan, each TV manufacturer copied the plan used on whatever set-top terminal he or she looked at. This led to considerable confusion on the part of the subscriber, who had to be trained that, in order to get such and such a program, he or she would have to tune to channel 37 (for example) with

---

* In fact, the TV tuners often had to have set-top terminals placed in front of them anyway. They were not well shielded against direct pickup. They didn't have adequate image rejection and were susceptible to overload from too many signals at the input. Further, the TVs often had inadequate adjacent channel rejection. These problems still required the use of a set-top terminal, to the considerable frustration of the subscriber, the store that sold him or her a "cable-ready" set, the TV manufacturer, and the cable company. Eventually, the FCC got involved.

one set top, channel 99 with another, or maybe channel 1 if he or she had a TV without a set-top terminal. The cable television and consumer electronics industries began discussing a standard band plan, under the auspices of the joint EIA (Electronic Industries Association) and NCTA (National Cable Television Association) engineering committee. The first plan, produced in the mid-1980s, was assigned the designation IS-6, an EIA designation for an interim standard. It covered allocations to 300 MHz. Even as an interim standard, it received widespread support from both industries. Set-top and TV manufacturers began applying it to new product. (However, even today there is product in use that predates this standard.)

Before IS-6 could be adopted formally, the industry began using frequencies above 300 MHz. The committee extended the plan to cover channels through 1 GHz and to define how channels above this would be allocated. The new standard was designated IS-132. The prototype standard was rather quickly produced, but a number of years elapsed before it was officially approved and sent to the American National Standards Institute (ANSI). The final EIA-assigned standard number was EIA-542. Later the standard was upgraded to show digital channel allocations, in the process becoming EIA/CEA-542 Rev. A. This standard is described in the remainder of this section. Some seemingly strange violations of a straightforward channel number allocation plan exist for historical reasons: a new standard must not cause undue problems for equipment already in use when it is adopted.

9.2.3 The EIA-542 Band Plan

Outline of the Plan

Figure 9.2 shows the relation between the standard cable channels of EIA-542-B and the FCC frequency assignments.[2] (Besides the standard assignments, EIA-542-B defines HRC and IRC frequency plans, shown in Figure 9.4.) The boxes in Figure 9.2 contain the FCC channel numbers for off-air assignments plus the traditional letter designations for cable-only channels. The new channel numbers as defined by EIA-542 are indicated above the boxes. Below the boxes are representative frequencies for the channel boundaries. Finally, the traditional cable band designations are shown below the frequencies. These band designations are not used very often today.

Channels 2–13 are identical to the FCC assignments for off-air broadcasting. The midband comprises channels 14 (121.25-MHz picture carrier) through 22 (169.25 MHz). Channel 23 (217.25 MHz) is the first channel above channel 13. From channel 23, the channel numbers are progressive through channel 94 (643.25 MHz). Channels 95–97 lie in the FM band, and channels 98–99 define the frequencies between the FM band and channel 14. Channel 100 picks up at 649.25 MHz. From there, the plan progresses one channel number per 6 MHz, up

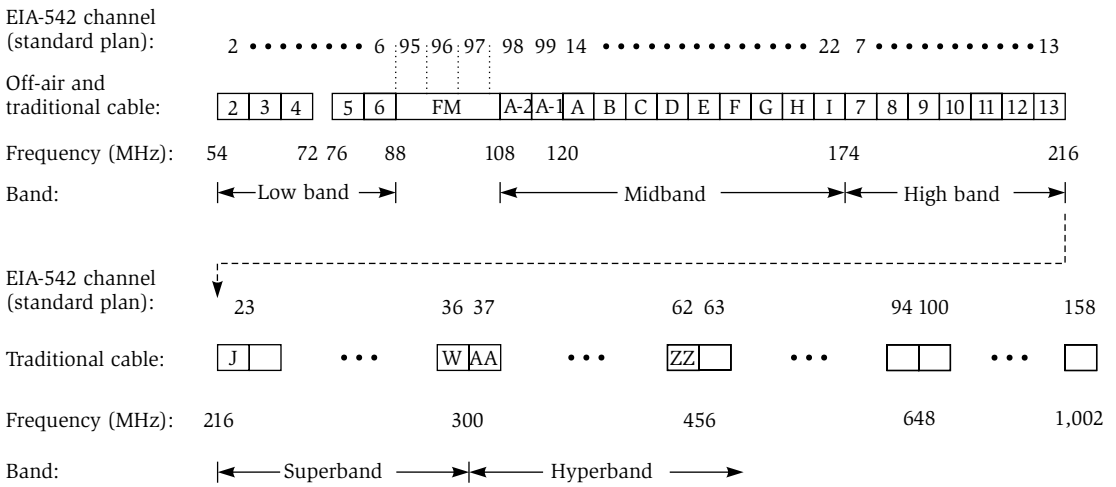

EIA-542 channel (standard plan):
2 6 95 96 97 98 99 14 22 7 13
Off-air and traditional cable:
2 3 4 5 6 FM A-2 A-1 A B C D E F G H I 7 8 9 10 11 12 13
Frequency (MHz):
54 72 76 88 108 120 174 216
Band:
Low band
Midband
High band
EIA-542 channel (standard plan):
23 36 37 62 63 94 100 158
Traditional cable:
J W AA ZZ
Frequency (MHz):
216 300 456 648 1,002
Band:
Superband
Hyperband

Figure 9.2 Relation between cable channels and FCC assignments.

to channel 158, 997.25 MHz, where the present revision of EIA-542 stops. Above this frequency, the plan specifies that the channel numbers shall continue to progress one channel number for each contiguous 6-MHz band.

Relation to the Off-Air UHF Band

Figure 9.3 shows the relation between the cable plan and FCC assignments in the UHF television band. The FCC plan is offset up 2 MHz from what we would expect if we were to keep adding 6 MHz to the VHF allocations. The cable plan fills the spectrum completely from above channel 13 to the UHF band. Consequently, the cable band edges are lower than the off-air band edges by 2 MHz.

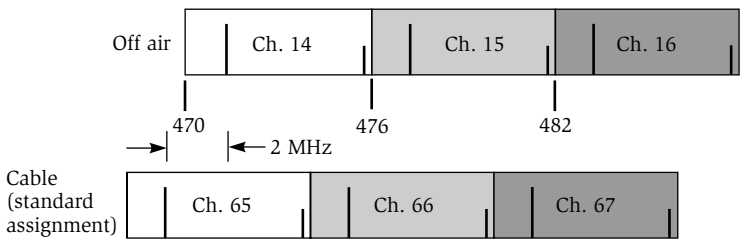


Figure 9.3 Relation between cable and FCC assignments in UHF band.

This can cause problems when these channels are used on cable systems. In the UHF region, the channel offset can result in a local broadcaster rendering two channels less desirable on the cable. For example, assume channel 15 was used off-air. The picture carrier from channel $15_{\text{off air}}$ will fall in channel 66_{cable}. The aural carrier from channel $15_{\text{off air}}$ will fall in channel 67_{cable}.

If the field strength of channel $15_{\text{off air}}$ is high, the sound carrier will produce beats in channel 67_{cable}. On the other hand, if the field strength of channel $15_{\text{off air}}$ is not too high, the offset can help cable reception. Because the picture carriers are offset by 2 MHz, the visibility of any beat produced is less than if the offset were not present. The so-called W curve, described in Chapter 2, shows that an interfering carrier offset from the picture carrier by 2 MHz is not as visible as if it was closer.

IRC and HRC Assignments

In order to control the effects of distortion in cable television distribution plant, some systems employ alternative frequency plans. Of a number of plans that have been tried over the years, only two alternative plans survive. Figure 9.4 contrasts these two plans with the off-air (standard) assignments. The low VHF band is shown since it is the most affected.

The incrementally related coherent (IRC) plan (lowest line of the figure) places all carriers, except for channels 5 and 6, on the same frequency as in the standard plan. All picture carriers are phase locked to a comb generator, whose output is derived from a common 6-MHz oscillator, but with the comb shifted 1.25 MHz high in order to place the picture carriers on their normal assignments. We can show that many third-order distortion products fall coherently on other

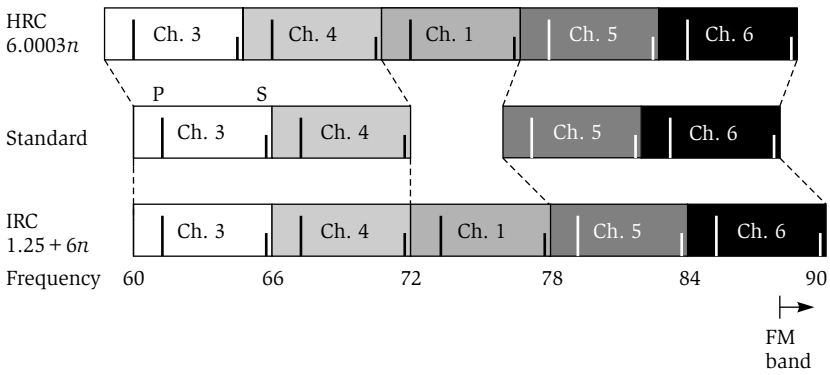


Figure 9.4 Relation between standard, IRC, and HRC plans.

carriers in this case, yielding lower beat visibility. The 4-MHz gap between channels 4 and 5 becomes 6 MHz because channels 5 and 6 must be shifted in order to accommodate IRC assignments. They are shifted up by 2 MHz, invading the FM band. This opens a 6-MHz gap between channels 4 and 5, which is assigned the designation "channel 1." (Some cable systems are using 48–54 MHz and calling it channel 1. This is not recognized in the plan.)

The other plan is the harmonically related coherent (HRC) plan (top row in the figure). All picture carrier frequencies are at harmonics of 6.0003 MHz. We explain the significance of the 6.0003-MHz frequency later. The picture carrier frequencies are derived by phase locking the picture carrier generators to a 6.0003-MHz comb. We can show that this causes all second- and third-order distortion products to zero beat other carriers, providing somewhat more immunity against distortion than does the IRC plan. All carriers are shifted low by 1.25 MHz, to fall on the harmonics of 6.0003 MHz. Again, the exceptions are channels 5 and 6, which are shifted high by 0.75 MHz. The combination of shifting channel 4 lower by 1.25 MHz, and channel 5 higher by 0.75 MHz, again creates a 6-MHz opening that is assigned the designation "channel 1."

Operation in the Aeronautical Bands

Cable systems wishing to put signals in the frequency bands utilized in the aeronautical service (108–136 and 225–400 MHz) are required to periodically measure leakage from their systems and to repair any leaks that exceed a certain standard. The standard is much more stringent than is the standard to which consumer electronics manufacturers are held. In addition to controlling leakage, signals carried on the cable system are required to be offset in frequency from the nominal aeronautical carriers. The nominal offset puts the cable picture carriers at the edges of the aeronautical channels. The offset requirements apply to any carrier that exists in the distribution plant at a level exceeding 10^{-4} watts (+38.75 dBmV), measured in a noise bandwidth of 25 kHz and averaged over 160 μs.[3] With realistic distribution plant signal levels, this often includes picture carriers, but rarely includes sound or data carriers.

Channel Assignments in the Aeronautical Band

Figure 9.5 shows the relation between cable carriers and aeronautical channels. The aeronautical bands are subdivided into aeronautical radio navigation and radio communications bands. The navigation channels are assigned channels in 50-kHz increments, and the communications bands in 25-kHz increments. The aeronautical carrier frequencies fall precisely on the nominal TV picture carrier frequencies. Because of this, cable systems must offset their picture carriers so that they fall on the aeronautical channel edges. In the aeronautical radio navigation bands, this necessitates a 25-kHz offset. In the communications band,

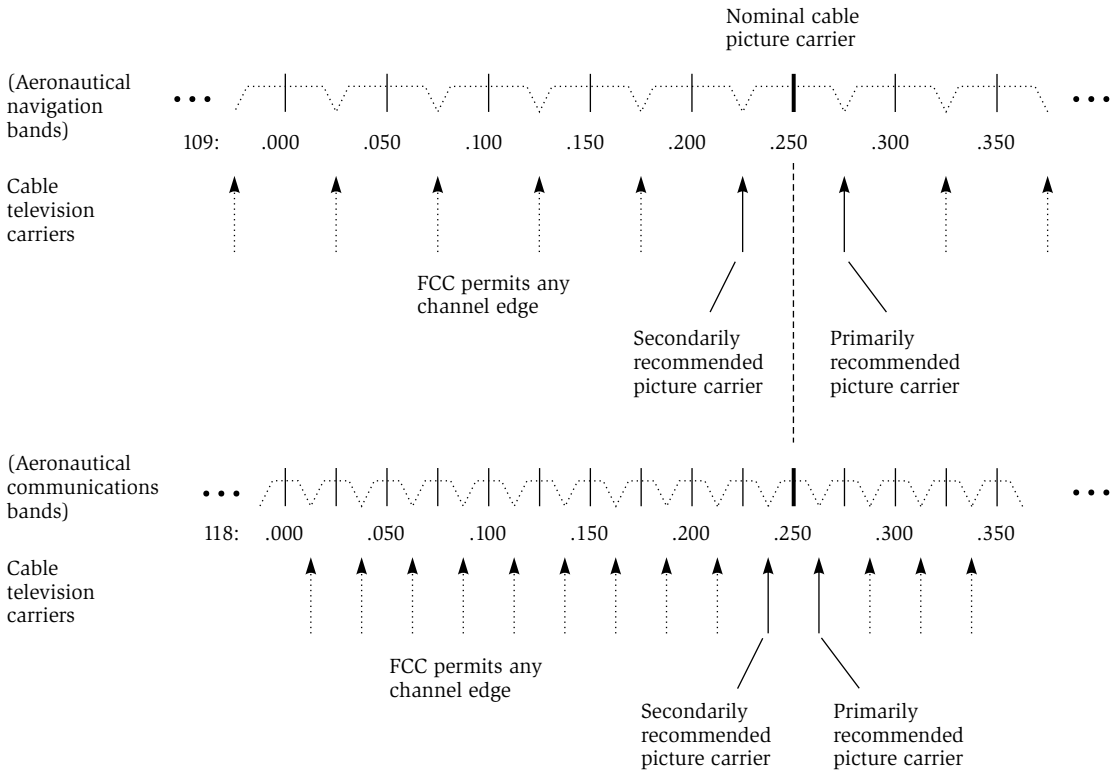


Figure 9.5 Relation between cable picture carrier frequency and aeronautical channels.

the offset is 12.5 kHz.* The FCC permits any aeronautical channel edge to be occupied by a picture carrier. In order to simplify the allocations, CEA-542-A or -B recommends picture carriers be placed at the next channel edge higher than the nominal frequency. This seemed to be the most common assignment used before the recommended standard was written.

Some cable systems, though, offset to the next lower channel edge, which is shown as the secondarily recommended picture carrier. One reason a system may want to use the negative offset is to help identify the source of leakage. Cable systems must formally measure leakage from their systems frequently, and must informally monitor leakage continuously. If two systems are adjacent

* For a number of years, there has been talk of splitting the aeronautical channels into two, halving the separation between center frequencies. If this ever happens, the now-required cable offsets will fall on carrier frequencies rather than on band edges.

to each other, it can be difficult to determine which one is responsible for a leak. When this condition exists, one of the systems will employ positive offsets, and the other negative offsets. Thus, by measuring the frequency of the leakage, it is possible to identify the system with the leak, allowing the correct repair crew to be dispatched. The recommended standard permits, but doesn't encourage, any offset up to ±250 kHz. It is felt that most TVs will accommodate this much offset without difficulty.

The tolerance on the carrier frequencies is specified by the FCC as ±5 kHz in the aeronautical band (elsewhere the tolerance is ±25 kHz). In order to accommodate HRC assignments within this limit, it is necessary to use a comb generator with a fundamental frequency of 6.0003 MHz and a frequency tolerance of ±1 Hz. Section 8.4.1 describes an HRC comb generator.

FM Band Usage

In some cable systems, spectrum is at such a premium that the operator has been driven to use the FM band for TV program distribution. This is not common, but is sometimes done. Also, some systems use the FM spectrum for digital music channels. The recommended standard permits the use of the FM band frequencies, but does not require that the TV tune these frequencies. The cable operator is advised that many "cable-ready" sets will not tune to these channels. Many TVs include traps in the FM band to reduce distortion effects in the tuner. These traps preclude reception of the FM band.

Alternative Modulation Methods

Acknowledging the use of digital modulation in 6-MHz channelization, the recommended standard proposes that any signal delivered to the consumer should fit in channels having the edges at the frequencies of the NTSC channel edges. This does not preclude distribution of programming in the plant using some other frequency plan, as long as any signal handed to the subscriber at RF follows the plan. For example, digital signals might be carried in a 12-MHz band on the cable. This would be permitted so long as that signal was never handed to the subscriber. If a device supplied by the cable operator converts the signal to something contained within one 6-MHz channel having the band edges shown in Figure 9.2, that system would comply with EIA-542-B.

9.3 Headend RF Management

Until recently, headends were truly broadcast-oriented structures. They were built to collect various signals, format them, and send the same set of signals to all subscribers. This has been changing. The transition of the industry from tree-and-branch networks to nodal networks has enabled reuse of frequencies on

different nodes. The industry has begun routing different signals to different nodes for telephony and data applications. Signals in these services are intended for individual subscribers rather than for all. In order to utilize bandwidth efficiently, it is desirable to determine the node location of an intended subscriber and to send signals bound for him or her to only that one node. This increases the efficiency of the cable network by allowing reuse of the same frequencies on each node, transmitting only information bound for someone in that node.

Besides the fairly obvious advantages of sending data and telephony signals to only the node requiring them, other services can benefit from this capability. For instance, the industry has talked about node-specific advertising, in which advertisements are selected based on the demographics of subscribers connected to that node. Yet another anticipated application is to supply customized video-on-demand (VOD) programming to individual subscribers.

RF management is the art of grouping signals bound for each node while maintaining level control in the most economical way possible. It involves physically segmenting the signals while maintaining isolation between them. It also involves handling return signals entering the headend, where it is desired to route signals to each of several return services, while minimizing the amount of hardware required.

9.3.1 Downstream RF Management

Figure 9.6 illustrates a segment of a modern headend. Signals are received from satellite earth stations and other sources. As shown, each signal received from the earth station may have a number of pieces of hardware associated with it. The satellite receiver, described in Section 8.7, converts the signal from RF to baseband NTSC, usually in association with a decoder to descramble the satellite signal. These two pieces of equipment are often combined into an integrated receiver-decoder (IRD) in modern headends. If stereo is offered on the channel, video from the IRD is routed through a stereo encoder (see Section 8.6) and then to the scrambler if used (see Chapter 21). Finally, the signal is modulated onto an RF channel in the modulator. A number of such signals are acquired and combined in headend combiner CM2. Additional signals, such as digital audio broadcasts, may also be included. The output of combiner CM2 is the total set of broadcast signals, which are supplied to all subscribers.

Besides these broadcast signals, node-specific services are becoming popular. These commonly include telephony, data services, and in the future, node-specific programming. These services are duplicated for each node; different signals are transmitted to each node, but all use the same set of frequencies. Adequate input ports must be provided for addition of future services. Unused ports must be terminated. Services for node 2 are combined in combiner CM5, and those for node 1 are combined in CM4. These signals are combined with the broadcast signals before being supplied to the individual node optica transmitters.

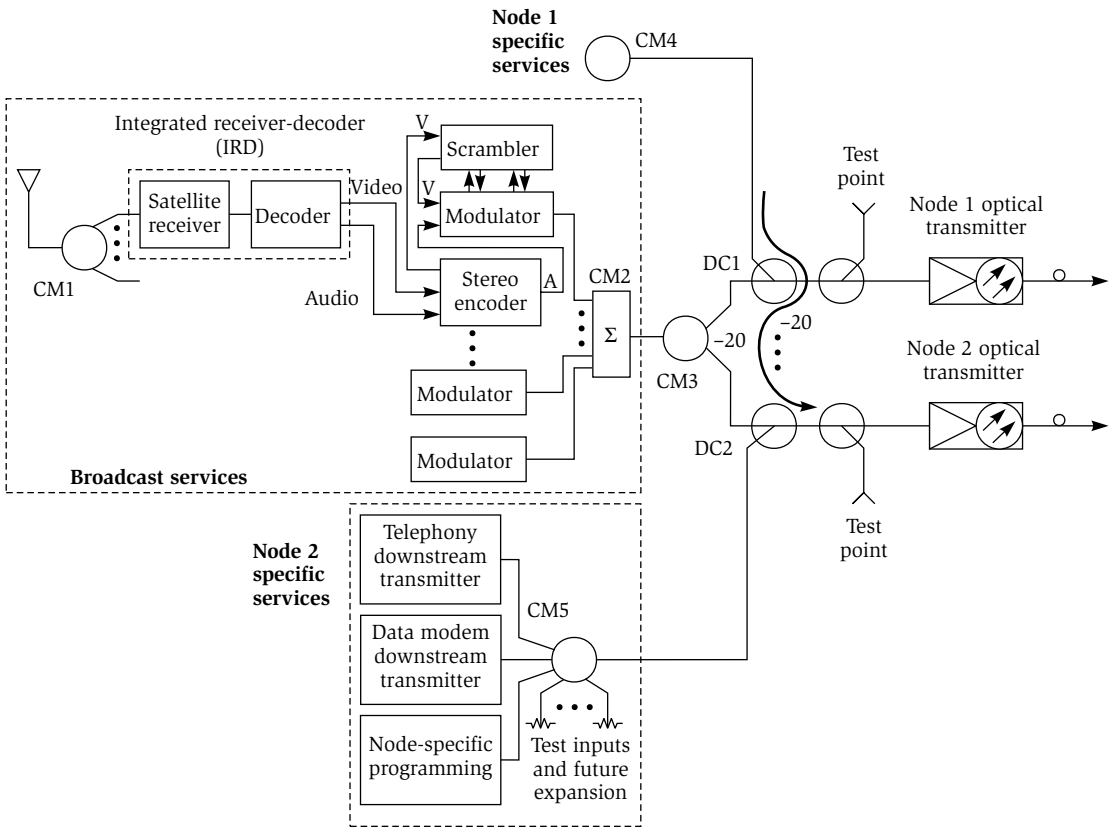


Figure 9.6 Downstream RF management.

The combining of broadcast and node-specific services is accomplished by first splitting the broadcast services in splitter CM3. One output of CM3 is supplied to directional coupler DC2, where signals bound for node 2 are coupled to the broadcast signals. Similarly, directional coupler DC1 is used to couple the broadcast signals with signals bound for node 1.

Level Management

A portion of the RF management issue involves physically arranging the combiners and directional couplers such that efficient use is made of the hardware, without causing problems for future expansion. Additional inputs are provided as illustrated at CM5 for services that may be added later.

Provision of additional inputs as shown does not come without penalty, however. Each time the number of ports provided at CM5 is doubled, the attenuation of signals from each existing service is increased by 3.5 to 4 dB. This means that at some point additional amplification may have to be provided.

Figure 9.7 illustrates a simple way in which the node 2 combining may be modified to reduce the level required of existing services, compared with Figure 9.6. In Figure 9.6, an eight-way splitter is provided at CM5, allowing eight services to be combined. The insertion loss of an eight-way combiner is approximately 12 dB (budgeting 4 dB per two-way combine). If this does not let the node 2 transmitters operate at comfortable levels, then Figure 9.7 can be used. Combiner CM5 is replaced with a four-way combiner, which has only 8-dB loss for each input. One of those inputs is from another four-way combiner CM6, which provides for additional services. By using two four-way combiners rather than one eight-way combiner, today's services can operate with 4 dB less level required at the respective specific services transmitters. The penalty is that future services will have to operate at levels 4 dB higher than they would otherwise (they see a 16-dB loss before directional coupler DC2), and one input is lost. Some cable operators consider this to be a good trade-off to allow the headend to be built simpler today — in some cases, using the Figure 9.7 configuration can save an amplifier in the combining network.

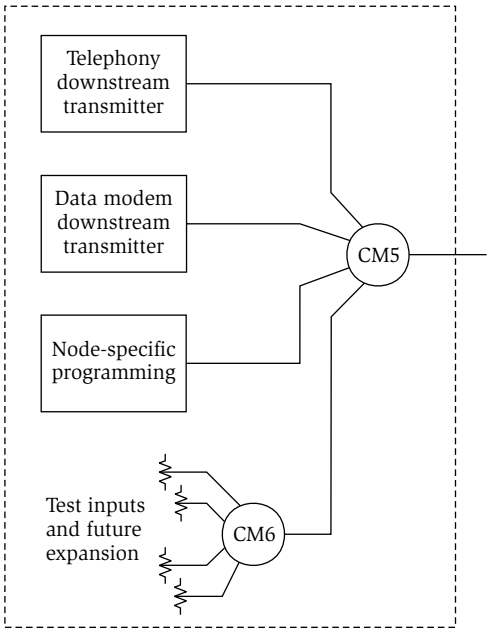


Figure 9.7 Alternative combining arrangement for node 2.

Signal Isolation

A major issue in both downstream and upstream RF management is isolation between nodes. This is illustrated in Figure 9.6, with the example of interference from node 1 into node 2. It is assumed that a similar service at the same frequency is used on each node. The problem is to avoid interference from node 1 into node 2, and so on. The heavy arrow of Figure 9.6 illustrates the example case. Signals from services specific to node 1 are coupled to the node 1 transmitter in directional coupler DC1. Theoretically, all the signal either is routed to the optical transmitter or is dissipated in the internal load resistor of DC1. If DC1 is a 17-dB coupler, then the power that appears at its output (connected to the node 1 optical transmitter), is −17 dB from the incoming signal.

In the real world, some power is coupled backward in the directional coupler and goes toward CM3. Theoretically, any power that impinges on CM3 from one of its outputs either is coupled to the input (connected to CM2) or is dissipated within CM3. In the real world though, some signal is coupled from one output of CM3 to the other. Thus, node 1 signal power from CM4 can wind up being combined with node 2 signal power, where it becomes an interfering signal.

It is assumed that the isolation of each directional coupler and combiner is 20 dB. In traversing the wrong way through DC1, the signal is attenuated 20 dB. In traversing from one output of CM3 to another, it is attenuated another 20 dB. Assuming then that it is combined with another signal of identical level in DC2, the signal from node 1 becomes a −40-dB interfering signal to a node 2 service.

The point at which interference becomes significant is a function of a number of factors. To provide a feel for the isolation necessary in headend combining, Table 9.1 lists some potential interfering situations and a recommended headend isolation number. You can see that as modulation density increases, more isolation is required. This is because higher-modulation density signals have less distance between modulation states and so can tolerate less interference (see Chapter 4).

The analog visual carrier recommended isolation is based on the traditional 60-dB threshold for interference, with 10 dB added to account for possible level differences between nodes. It is recommended that, even if this much isolation is provided, the two video sources be phase locked since this has been shown to significantly reduce the visibility of interference. Phase locking will afford additional

Table 9.1 Recommended headend isolation

Interference	*Recommended Isolation (dB)*
Analog visual to analog visual	70 dB (phase lock recommended)
QPSK to QPSK	33 dB
64-QAM to 64-QAM	54 dB
256-QAM to 256-QAM	61 dB

margin to the threshold of visibility. It is known that the threshold of visibility can be more severe than −60 dB under optimum conditions, but this number, derived from work done at Bell Labs in the 1950s, is taken as a working number.

The recommended isolation for digital signals is derived from the crossing of a decision state threshold, taking into account the peak-to-average ratio of an unfiltered signal. No allowance is made for demodulation errors or filtering, but a 10-dB allowance is made for level difference, and a 20-dB allowance is made because other sources of signal degradation exist. The headend should not be a limiting factor.

9.3.2 Upstream RF Management

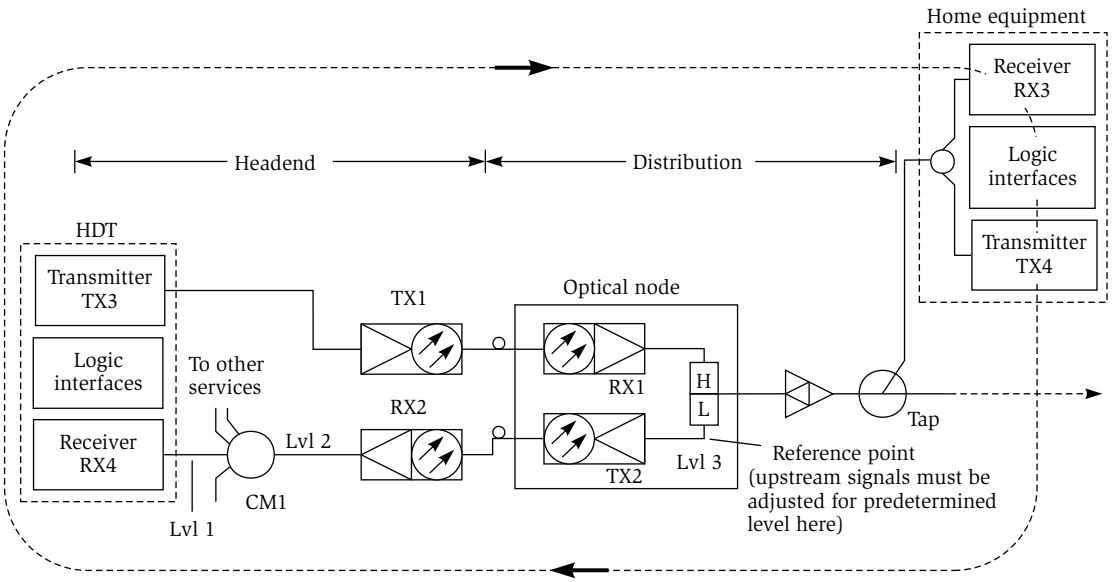
Some of the issues in RF management in the upstream direction are similar to those in the downstream direction, but upstream management introduces some unique issues. Signal isolation issues are similar to those in the downstream direction, but tend to be more complex owing to the need to share resources. Level issues become more complex because changes in the headend can easily affect operation of return components in the field, with operating personnel being unaware of the consequences of changes they make.

Level Management

Level management in the upstream direction is complicated because of the need to maintain accurate levels at the node (in the field) using long loop automatic level control (ALC). This issue doesn't have a close analog in the downstream direction and is not familiar to a lot of practitioners. Figure 9.8 illustrates a long loop ALC. The host digital terminal (HDT), a generic name for the headend equipment used with any two-way service, consists of a receiver for the upstream signals, a transmitter for the downstream signals, and logic to control those signals, as well as all the components required to operate the appropriate service. The topic of upstream level management is covered more fully in Chapter 16. The purpose of including this information here is to emphasize how the headend must be configured to allow it to function smoothly.

Downstream signals are transported to the optical node, where they are transferred to the coaxial network and to the home receiver, RX3. In the upstream direction, the home transmitter, TX4, supplies upstream signals, which are transported to the optical node and then supplied to the return optical transmitter, TX2. From TX2, the signal is transported to the headend receiver RX2, then supplied to the RF receiver RX4 in the HDT.

Each HDT using the return path should establish a long loop ALC as shown, in which the return transmitter (TX4 at a subscriber's home) is commanded to increase or decrease its level until the level returned to the headend subsystem, at Lvl 1, is some target value. After the return system is aligned, it is imperative that the gain between the headend optical receiver, RX2, and the HDT not change. If a new service is added, then power will have to be tapped from optical receiver



An ALC loop is set up between the headend and each home system individually. The headend senses the level of the return signal and adjusts the upstream transmitter such that the correct level is received at the headend receiver, Lvl 1. The most critical level is that at the reverse optical transmitter, Lvl 3.

Figure 9.8 Long loop ALC control of return transmission.

RX2 to supply the receiver of the new service. If this has not been accounted for initially, by provision of extra ports at CM1, the ports added will change the gain between Lvl 2 and Lvl 1. The existing subsystems will respond by increasing level at the home transmitter until the level at Lvl 1 has returned to its previous value. This will increase the level at Lvl 3, potentially overdriving the return optical transmitter, TX2. To avoid this, it is imperative that an adequate number of reverse ports be provided at CM1 when the headend is built.

Signal Isolation

Signal isolation issues in the reverse direction tend to be more difficult than in the forward (downstream) direction because many operators combine signals from different nodes for some services and not for others. For example, a telephony service may employ a separate HDT for each node. This would be done to permit good availability of service, with little chance of blockage. A modem service may combine two or more nodes to a common HDT to achieve better utilization of equipment — modem systems tend to degrade gracefully as more people use the service. Finally, a status-monitoring system doesn't need

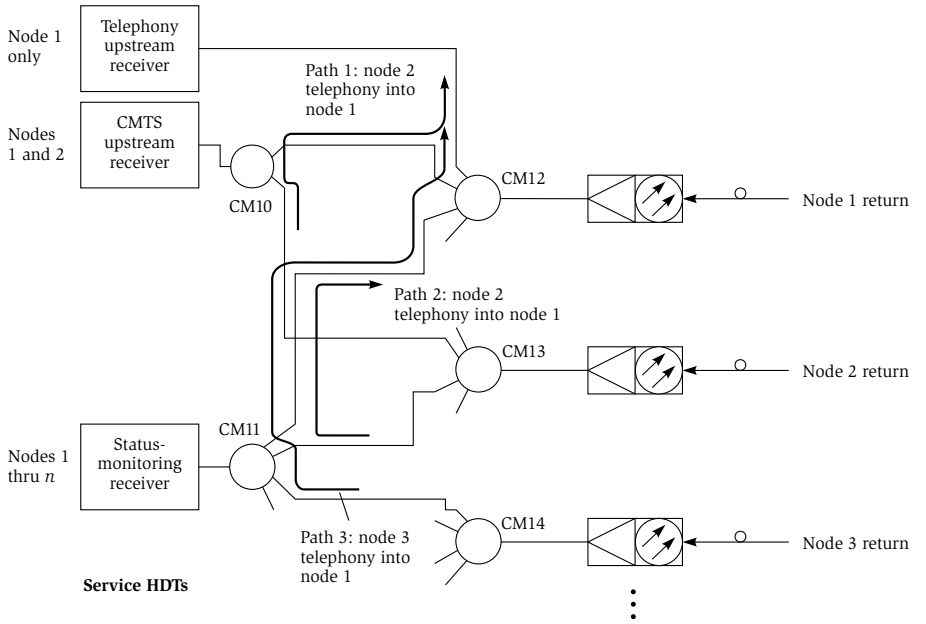


Figure 9.9 Return path signal coupling.

constant communications, so it is possible to combine signals from many nodes to minimize the number of status-monitoring HDTs required.

Figure 9.9 illustrates such a situation. The telephony system has a dedicated HDT for node 1. A data CMTS is shared between nodes 1 and 2. Note that the CMTS *may* serve two nodes downstream but have individual upstream receivers. This helps compensate for the limited upstream bandwidth. Finally, a status-monitoring HDT is shared among many different nodes, of which three are shown. Splitter CM12 divides signal from node 1 to the telephony subsystem, the data subsystem, and the status-monitoring subsystem, and it has one or more additional ports for other services. As a minimum, a test point is needed, and ideally, the operator has made provision for future service expansion. Splitter CM13 provides the same service for node 2, as does CM14 for node 3.

Combiner CM10 has two inputs, from nodes 1 and 2, to allow them to share the same data modem HDT. Combiner CM11 has many inputs to allow many nodes to share the same status-monitoring HDT. All this sharing promotes efficient equipment utilization, but works against performance in that isolation is a problem. It is assumed that like equipment uses the same return frequencies on all nodes. This is usually the case, due to the scarcity of useful return spectrum.

In this limited example, three sneak paths exist to couple telephony signal power from nodes 2 and 3 to the node 1 HDT. Paths 1 and 2 both couple power

from node 2 to the node 1 HDT. Path 1 couples power from node 2 via CM10 and CM12. Path 2 couples power from node 2 to node 1 via CM11 and CM12. Path 3 couples power from node 3 to node 1 via CM11 and CM12. The power arriving at the node 1 telephony HDT adds from each sneak path.

In the example, if all combiners exhibit isolation of 20 dB, and all nodes have equal level return signals, then the total relative interference seen at the node 1 telephony HDT is three times the $10^{-40/10} = 10^{-4}$ provided by one sneak path. That is, each interference path traverses two combiners or splitters, each of which exhibits 20-dB isolation, so each interfering signal is −40 dB from the desired signal, and there are three such signals in the example. This works out to an interference level of $10 \log (3 \times 10^{-4}) = -35.3$-dB interference from the three sneak paths, assuming equal signal levels. In order to get back to parity with one sneak path, which would exist with no status-monitoring HDT sharing, the isolation of each combiner must improve by 4.7 dB. We can see the problem if the sharing of the status-monitoring HDT is spread over many nodes.

Isolation between nodes can be improved in several ways. Some operators characterize each combiner in the frequency band of interest and choose ports having the best isolation. This works for a limited number of ports in use, but as ports are filled up, the technique will ultimately break down. Sometimes attenuators in the combiner inputs can be used to effect improvement in isolation at the expense of more signal loss.

Filters can be used. For instance, suppose the status-monitoring HDT used a fixed return frequency of 8 MHz. If each input to CM11 had an 8-MHz bandpass filter, then many of the sneak paths used by the telephony subsystem (which probably uses frequencies above 20 MHz), would be broken. The problem, though, is that filters can limit frequency sharing between different services.

Finally, amplifiers may be used to effect isolation. A low-gain, high-isolation amplifier in each input to CM11 will significantly reduce the leakage. This, of course, is at the expense of more cost, possibly more distortion (though this can usually be controlled), and maybe a slight reduction of availability of the subsystem. An amplifier does not have unlimited isolation: many have as little as 15-dB isolation, but judicious selection of an amplifier and padding can often make the project succeed. Many operators use amplifiers as a last resort; nonetheless, the technique is gaining in popularity as headend RF management becomes more complex.

As a final note, it would be desirable to use some attenuation on each splitter or combiner output that goes to any receiver. Isolation of each splitter/combiner is enhanced by so doing. Power coming into one leg of CM10, for example, at the telephony receive frequency will probably see a poor return loss at the data modem upstream receiver. When the signal encounters a poor return loss, it must be reflected to the combiner, and one-half the reflected signal will be routed to the telephony upstream receiver. This mechanism reduces the isolation of a splitter compared to that isolation of which it is capable. An attenuator on each output port will improve the situation.

9.4 Headend Fiber-Optics Management

Almost all modern headends drive some amount of fiber-optic cable. The use of fiber-optic transmission has enabled many modern service offerings. However, fiber optics has its own set of requirements and limitations. Fiber-optic components and handling techniques have evolved rapidly. Chapter 12 describes the three methods of joining fiber-optic cables: fusion splicing, mechanical splicing, and remateable connectors. Optical connectors may be installed directly in the field, but it is much better to use factory-manufactured pigtails, which are cut and each end fusion spliced onto the incoming optical cables. Installation of optical connectors traditionally has required difficult polishing best left to a factory. However, some newer connector systems now allow connectors to be put on in the field.

9.4.1 Fiber Distribution

Though many techniques are used to bring fiber cable into headends, a popular method is shown in Figure 9.10.[4] Outside plant fiber cables are brought to the headend at the right of the figure. The jumpers are fusion spliced in the outside cable entrance facility (OCEF), which typically is located just outside a headend wall. A pigtail brings signals into the fiber distribution bay inside the headend. The pigtails are often made by cutting a factory-manufactured jumper in two, yielding two pigtails. The jumper can be tested before it is cut; then two known good pigtails are produced when it is cut. Each pigtail is fusion spliced onto an outside fiber in the OCEF.

Connectors on the pigtails are permanently mounted to the rear of optical couplers (bulkhead connectors) in the fiber distribution bay. A fiber jumper is plugged into each active cable and couples signals to and from the electronic equipment racks, where the optical transmitters and receivers are located. Optical transmitters and receivers are usually supplied with a pigtail that couples power from the laser diode or to the receiver diode. The pigtail normally terminates on a coupler (bulkhead) located on the chassis of the transmitter or receiver. The jumper from the fiber distribution bay plugs into that pigtail.

Of course, one could simply bring the outside plant cable into the headend, directly to the transmitters and receivers, bypassing the fiber distribution bay. The advantage of the fiber distribution bay and optical jumper is that they allow rapid recovery from plant damage and equipment failure with minimum sparing. The remateable connection at the transmitter or receiver is useful for recovering from equipment failure: if a transmitter or receiver fails, it can be replaced and the cable plugged into the new device. If a fiber fails, the connector at the fiber distribution bay permits rapid patching to a spare fiber strand.

Typically, the same bundle of fiber strands (individual optical conductors) will service several nodes, as suggested in Figure 9.10. Spares will be carried to each

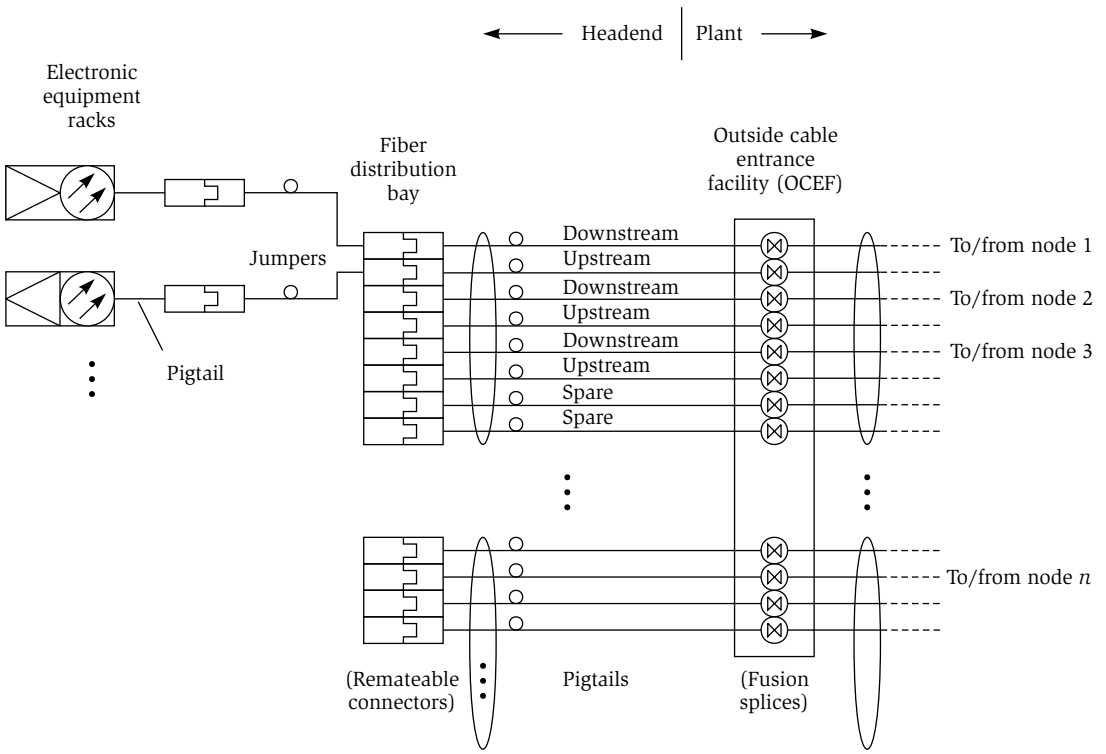


Figure 9.10 Fiber distribution in headend.

node. However, often it is desired that a spare optical strand will serve either a transmitter or receiver, depending on what fails. In some headend configurations, the transmitter and receiver servicing a particular node are not located close to each other. This means that it is not feasible to bring a particular strand up to, for example, a transmitter. If the strand were later needed for a receiver, that strand would have to be extracted from typically very large cable bundles in the headend and routed to the receiver. This would eventually destroy the organization of wiring in the headend, reduce reliability because of cable damage, and be time consuming. With the fiber jumper as shown, it is a simple task to move the appropriate jumper to the failed equipment on one end and to the appropriate outside plant strand.

The fiber distribution bay provides a convenient place to perform optical splitting when used. Optical splitting is used to route the output of one downstream transmitter to more than one node, and also to route light from one transmitter in two directions in a ring configuration. Ring configurations are used to provide path redundancy: if one path is damaged, such as from an overhead fiber being snagged, the second path takes over until the fiber is repaired.

In addition to covering the failure, use of the fiber distribution bay permits more rapid deployment of additional services that employ the unused ("dark") fiber running to a node. We do not have to worry about locating new equipment where the fiber is physically terminated; rather, a new jumper can be routed in the headend as desired.

It is also possible to omit the OCEF, bringing the outside cable directly to the fiber distribution bay. The outside cable will still need to be fusion spliced to a pigtail. Besides providing a convenient place to terminate outside plant fibers (if not done in an OCEF), the fiber distribution bay provides for labeling of the fibers. It also provides protection to the fiber, by shielding it from accidental contact, and provides for adequate strain relief and acceptable minimum bending radius.

9.5 Signal Quality Tests

In this section, we illustrate some simple yet effective basic quality tests that should be performed on video and audio signals in the headend. Also see Chapter 2, which contains information on the baseband video signal. Quality measurements for digital signals are covered in Chapter 4.

9.5.1 Visual Modulation Definition

Figure 9.11 shows a segment of the same video waveform used in Chapter 22 for illustration. The waveform is amplitude modulated onto an RF (or IF) carrier to the standard depth of modulation prescribed for the standard in use. As

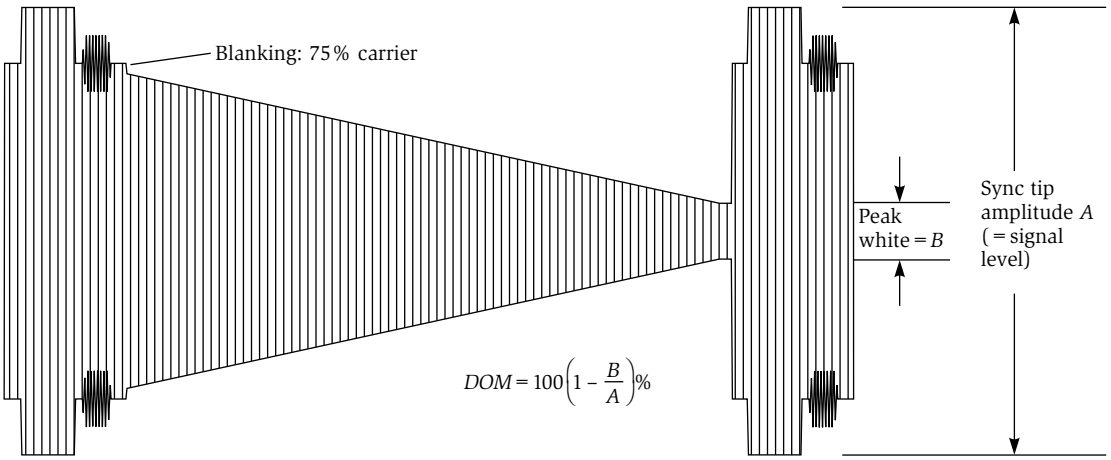


Figure 9.11 Definition of the modulated visual signal.

described in Chapter 8, the waveform is clamped on sync tips, such that the peak amplitude of the modulated waveform, which occurs during sync tips, is constant regardless of the modulation. The amplitude, A, during the sync tip is defined as *the amplitude* of the signal. As the video goes toward white, the instantaneous amplitude of the carrier drops. By definition, the peak white video modulates the carrier to 12.5% of the sync tip amplitude (that is, $B = 0.125A$, NTSC system). Video modulation is characterized by *depth of modulation* (DOM), which expresses the percentage of the distance the carrier transitions from peak sync amplitude toward zero amplitude:

$$DOM = 100\left(1 - \frac{B}{A}\right)\% \tag{9.1}$$

In most PAL systems, the depth of modulation is limited to 80%, whereas in NTSC, a depth of 87.5% is standard. The carrier is never totally cut off. This is important for several reasons. Many envelope detector circuits, as the envelope drops, tend to degrade in terms of their linear detection of the modulated signal. This is due to the diode curve, which exhibits a square law characteristic at low bias levels. This would tend to translate into differential gain (and differential phase due to capacitance modulation) if the carrier amplitude drops too far.

Another reason to maintain some carrier is the almost universal use of intercarrier sound detection in consumer receivers. The sound carrier at intermediate frequency is mixed with the picture carrier to produce the so-called *intercarrier* frequency, the difference between the two. Use of intercarrier detection is important to reduce the effects of incidental frequency modulation of the sound carrier. It also provides a convenient way to place the sound carrier in the middle of a passband that is narrow with respect to the accuracy of frequency conversions preceding it. However, if the picture carrier goes too low in level, the intercarrier detector will interrupt and/or phase modulate the sound intercarrier, with resultant buzz in the audio.

Thus, by standard and by FCC regulation, the depth of modulation of the visual signal is limited to 87.5%. Note that the color subcarrier is not included in the depth of modulation issue: often depth of modulation is measured using a low-pass filter that eliminates the color subcarrier from the measurement. The issue of whether to include the color subcarrier in the measurement has long been controversial. For most programming, the issue is somewhat academic since most bright (high-luminance) picture elements tend to have little color saturation, so little color subcarrier is present near peak white. (A notable exception, and one that has caused some headache, is the famous Big Bird of PBS's *Sesame Street*. His bright yellow color has been known to cause clipping and out-of-channel emissions from older transmitters.)

If color information does cause overmodulation, there may be some color and luminance distortion, but it is in a portion of the color palette where the distortion tends to be somewhat less visible. Intercarrier problems in the audio

tend to be masked because the color subcarrier is of such a high frequency compared with audio.

Measuring Depth of Modulation

Figure 9.12 shows a general setup for measuring depth of modulation (DOM). The measurement can be made in service using a test point on the modulator or an external directional coupler. Two DOM measurement methods are shown. A spectrum analyzer can be used as a linear receiver. Alternatively, a demodulator with a zero chopper can be used with an oscilloscope or waveform monitor. The waveform monitor is a specialized oscilloscope that is calibrated in IRE units and in depth of modulation.

DOM Measurement with Spectrum Analyzer

A spectrum analyzer may be used by tuning it to the picture carrier. Suggested settings are shown in Figures 9.12 and 9.13. In order to measure DOM, two references must be established. Sync tips form one reference, and are set eight divisions above the bottom of the analyzer screen. The other reference is zero carrier, which by the nature of the analyzer display is on the bottom of the screen. Note that the video waveform is presented upside down from the way it is normally drawn: the analyzer displays increasing signal level toward the top of the screen. Recall that sync represents the highest amplitude of the signal. Also, note that most spectrum analyzers do not have a wide enough bandwidth setting to pass the color subcarrier, so the measurement won't include color signals.

DOM Measurement with Precision Demodulator and Oscilloscope

Figure 9.14 illustrates the waveform obtained with the lower configuration of Figure 9.12. As we noted, two points of reference are needed in order to establish

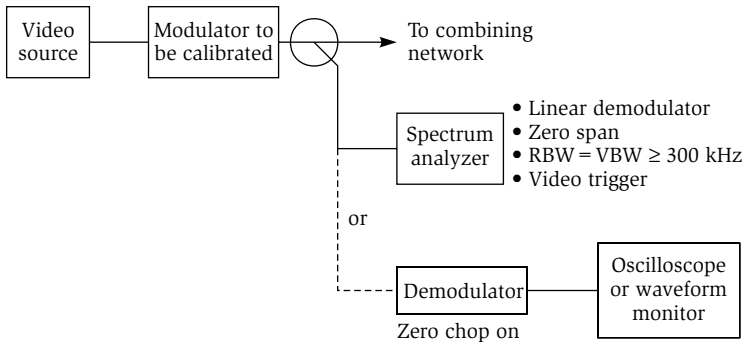


Figure 9.12 Measuring DOM.

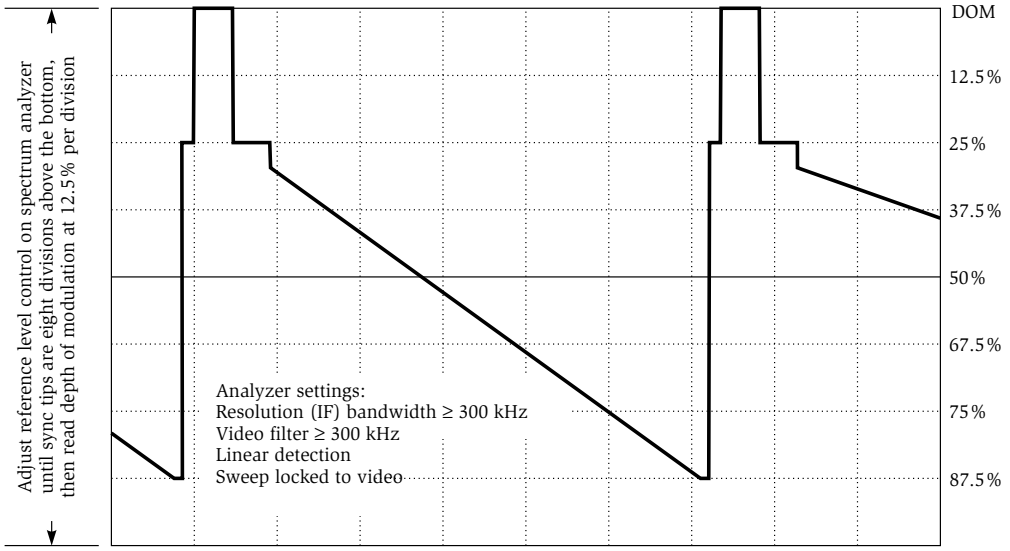


Figure 9.13 DOM display using a spectrum analyzer.

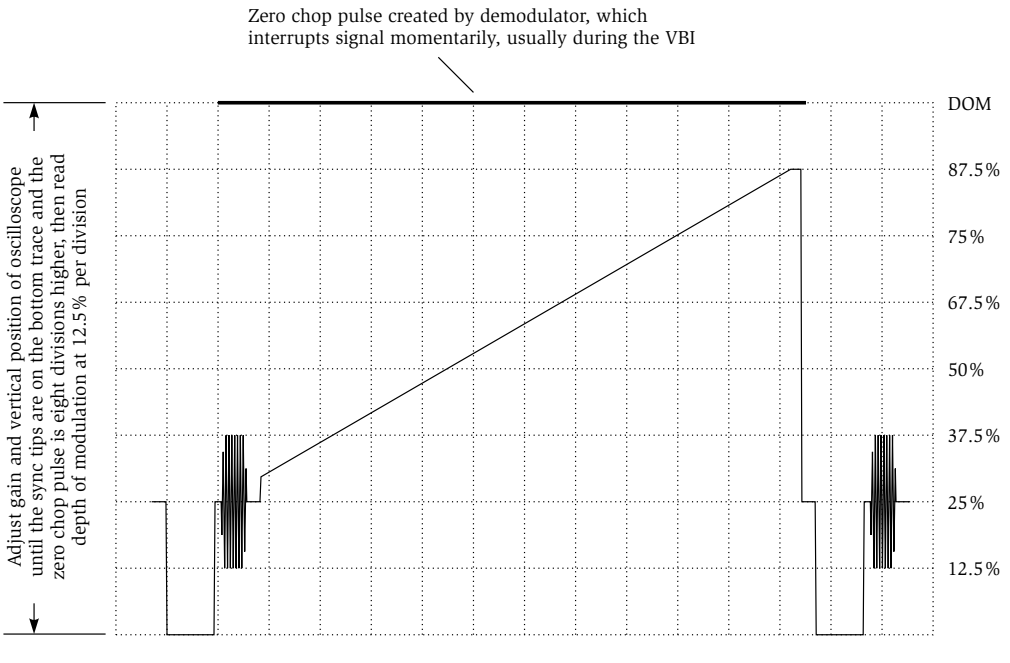

Zero chop pulse created by demodulator, which interrupts signal momentarily, usually during the VBI
Adjust gain and vertical position of oscilloscope until the sync tips are on the bottom trace and the zero chop pulse is eight divisions higher, then read depth of modulation at 12.5% per division
DOM
87.5%
75%
67.5%
50%
37.5%
25%
12.5%

Figure 9.14 DOM display using a precision demod and oscilloscope.

a scale for measuring DOM. The sync tips at the bottom of the waveform are one reference. The other must be related to the white level. A demodulator can insert such a reference by interrupting the signal before the detector. Usually, the interruption is during the vertical blanking interval. This "zero chop pulse" is placed eight divisions above the sync tip. Then peak depth of modulation is adjusted until peak white is seven divisions above sync. Zero chop pulse generation is described in Chapter 8.

9.5.2 Aural Modulation

Equal in importance with setting visual modulation is the setting of aural modulation. One of the most frequent complaints against cable systems is inconsistent audio level from one channel to another and from one program or commercial to another. This complaint has been leveled against broadcast stations as well. An objective measurement of audio level is difficult if not impossible owing to the subjective nature of human perception of sound. Much research has been done in this area, and a number of methods exist for making such measurements. This is not intended to be a thorough treatment of the subject. Rather, it is intended to impart some concepts that should prove useful in monitoring the audio being transmitted.

Deviation of the Sound Carrier

Television sound is frequency modulated onto the sound carrier in both NTSC and PAL television systems (it is amplitude modulated in SECAM). In frequency modulated systems, the amplitude is constant, but the frequency changes with respect to the center, or rest, frequency. The output of the detector is proportional to the deviation with respect to that center frequency. Note that the amplitude of the sound carrier is not a factor in loudness of the sound. The more the carrier is deviated, the louder the resulting signal. The art of setting the perceived loudness of the sound is often called "setting the deviation."

It is assumed that some facility for measuring deviation of the aural signal is available. This facility may be a meter in a modulator or a stereo encoder. It may also be an external demodulator with, or connected to, some sort of metering means. It may be a light that flashes at the prescribed peak deviation. This is undesirable as *the* modulation monitor because it gives so little information. However, many operators use it because of its low cost. The lights were intended to be overdeviation warning indicators only but have been used for setting deviation.

Confirming Accuracy of the Deviation Indicator

The modulation indicator must be properly calibrated. A particularly accurate method to ensure this, and one used in many labs, is the Bessel null technique. It

utilizes the fact that, when frequency modulating a carrier with a sine wave, the carrier component goes to zero (all the power is contained in the sidebands) when the modulation index, M, is 2.4. This point is often called *carrier null*. The modulation index is defined as the ratio of peak deviation to modulating frequency:

$$M = \frac{\Delta f}{f_m} \tag{9.2}$$

In order to null the carrier component, we deviate the carrier by 25 kHz (the prescribed peak deviation for NTSC sound) using a modulating frequency that makes $M = 2.4$:

$$M = \frac{\Delta f}{f_m} = \frac{25\,\text{kHz}}{10.42\,\text{kHz}} = 2.4 \tag{9.3}$$

The frequency of 10.42 kHz for the modulating signal is chosen to make the equation correct. When we modulate with a 10.24-kHz sine wave and find the first carrier null, we know that the deviation must be 25 kHz.

Carrier null is easily detected using a spectrum analyzer as a monitoring instrument and a sine wave generator as a source. Set the analyzer to distinguish between sidebands and the carrier. Set the oscillator frequency to 10.42 kHz. Begin increasing the amplitude of the oscillator output, starting from near zero, while watching the carrier component of the analyzer display. When the carrier component (not the sidebands) first goes to zero, the modulation index is 2.4, and the deviation should read 25 kHz on the indicating instrument. (Note that the carrier will go through zero for certain higher-modulation indices also.)

Exercise caution when making this measurement so as to properly account for preemphasis. The aural signal is preemphasized, as explained in Chapter 8, to reduce noise. At a modulating frequency of 10.42 kHz, preemphasis adds +14 dB to the low-frequency deviation. If the indicating instrument reads the nonpreemphasized signal, then it will read 14 dB low in the test described.

In countries using 50-kHz peak deviation and 50-μs preemphasis, the same procedure may be used, except that, at the indicated test point, the deviation should read −6 dB from peak (because the standard peak deviation is 50 kHz rather than 25 kHz). The correction for preemphasis, if needed, is 10.7 dB.

Often the actual modulation level used is backed off from peak by perhaps 6–7 dB if a traditional VU type meter ballistic is used so as to account for the dynamic response of such a meter. If a peak indicating instrument is used, this backoff is not needed.[5] Unfortunately, some meters that have been provided on cable modulators don't have dynamics that correspond to any particular standard. A meter that is accurate for a continuous sine wave may have significant error when measuring real audio.[6]

Routine Audio Monitoring

Figure 9.15 illustrates the setup for detecting the Bessel null using a sine wave audio oscillator and spectrum analyzer (switch down). This is the most accurate way to calibrate an audio deviation indicator. The figure also shows an audio demodulator and stereo decoder connected to an *xy* oscilloscope (switch up). This is a particularly convenient way to routinely monitor audio quality. Some broadcasters have such a monitoring facility permanently installed in their audio signal chain. An experienced engineer can obtain a lot of information quickly from such a display. One audio channel (the left as illustrated) is connected to the *x* input of the oscilloscope, and the other is connected to the *y* input. The two channels must be calibrated for identical gain, and the position controls must be adjusted for zero crossing at the center of the screen.

Figure 9.16 illustrates a number of situations that can be detected using this pattern. Figure 9.16(a) illustrates a properly calibrated setup with a monaural signal. Since each channel of a mono signal contains the same voltage, the display should be simply a straight line from lower left to upper right.

Figure 9.16(b) illustrates a mono signal, but with the left and right channels out of phase. (An out-of-phase signal cannot be properly located by the human ear. When listened to on a mono TV, there will be no sound.) The length of the line is related to the peak deviation of the signal. Because a signal normally spends only a small fraction of the time at peak deviation, the line will become dimmer the farther you look from the center of the screen. Figure 9.16(b) also illustrates conventional quadrant designations: the majority of a proper in-phase signal will reside in quadrants I and III. An out-of-phase signal will largely reside in quadrants II and IV.

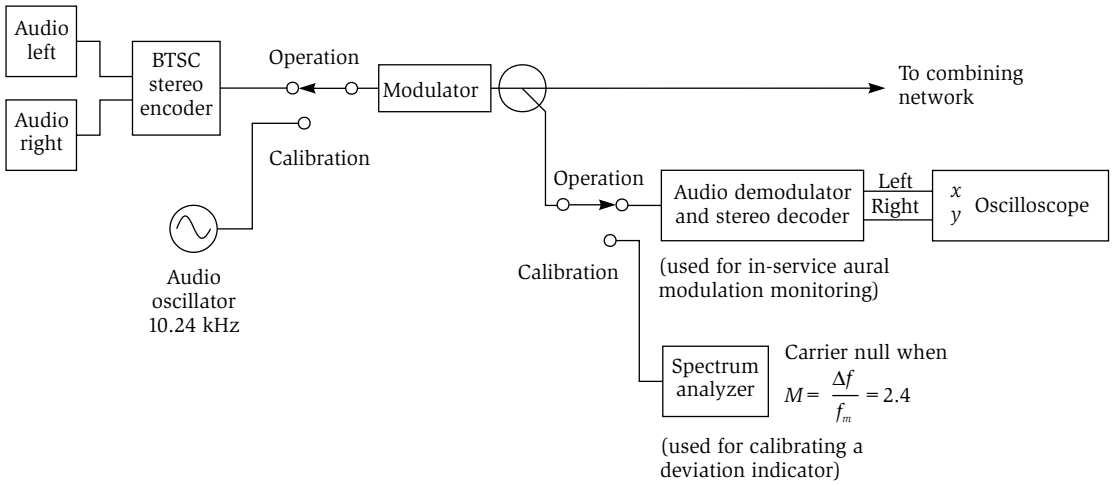


Figure 9.15 Audio monitoring options.

Figures 9.16(c) and (d) illustrate a stereo signal. The loci of all points occupied is "smeared" because the same signal is not being applied to each channel. The more the signal deviates from the mono axis of Figure 9.16(a), the greater the channel separation. (Compare Figure 9.16(c) with Figure 9.16(e), which shows less separation.) Low or high separation in program material is often a fact, not necessarily an error condition. Once an engineer becomes

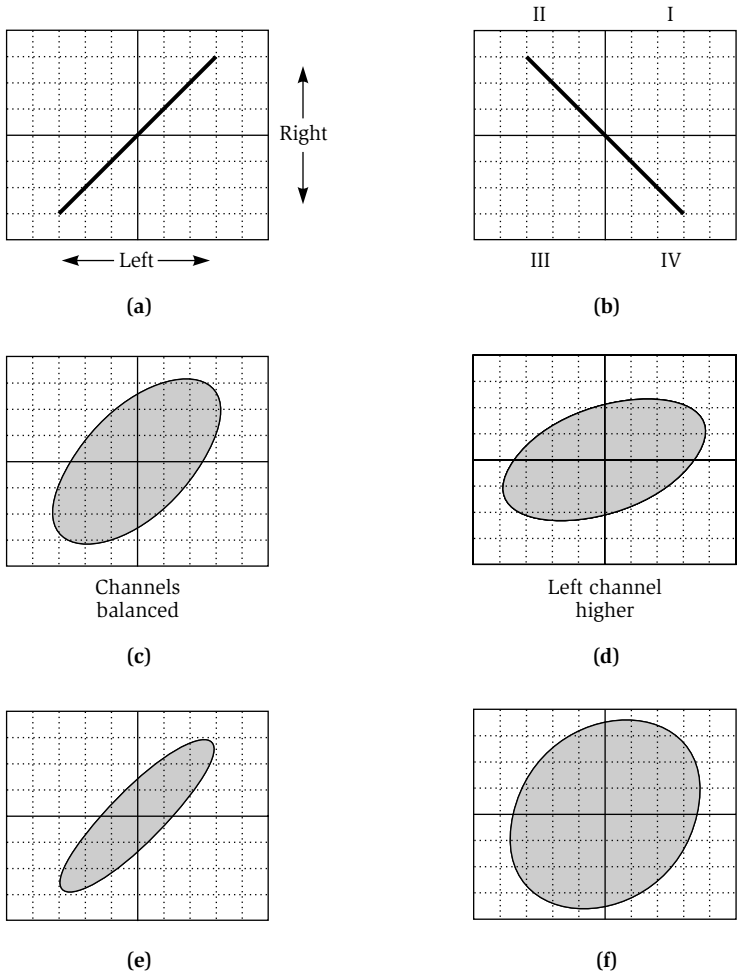


Figure 9.16 Oscilloscope patterns for stereo audio. (a) Mono, in phase. (b) Mono, out of phase. (c) Stereo, in phase, higher separation. (d) Same as (c), except left channel high. (e) Stereo, in phase, lower separation. (f) Synthesized stereo may show more occupancy of out-of-phase quadrants.

proficient at looking at these waveforms, he or she knows whether the separation is normal for a particular type of program material.

If the balance is not correct, then the waveform will be "tilted." The tilt will be down if the left-channel level is higher, as illustrated in Figure 9.16(d). If the right-channel level is higher, the waveform will be tilted up.

Notice that, though the majority of a stereo trace appears in quadrants I and III, some points appear in quadrants II and IV. This is to be expected with a stereo signal. Figure 9.16(f) illustrates a synthesized signal. Depending on the method of synthesis, more power may occupy out-of-phase quadrants II and IV. Some programmers use synthesizers when they are carrying programming that originates in mono.

9.5.3 Measuring Picture and Sound Carrier Levels

No measurements are as basic to setting up a headend as are measuring the picture carrier level and the sound carrier level. This may be done with a special-purpose signal level meter, or it may be done with a spectrum analyzer. The spectrum analyzer yields more information for the operator who understands it, but it is more expensive, more difficult to use, and capable of providing misleading results when used improperly. The signal level meter makes the reading without requiring setup, but it can be more difficult to use if many channels must be compared, and it gives less information.

Measuring Picture Carrier Level with a Spectrum Analyzer

Several points must be considered. The IF (resolution) bandwidth must be set wide enough to admit enough of the sync tip sidebands to allow the trace to reach its maximum. The video filter must also be set wide enough to allow the detected signal to reach full amplitude. As a practical matter, this means that both the IF and video bandwidths must be set to at least 300 kHz. The IF bandwidth must not be set so high that adjacent channel energy affects the measurement. Also, the sweep speed must be set low enough that you will see a number of horizontal lines on the waveform. With the correct setup, a spectrum analyzer can yield excellent results, though.

Scrambled signal level cannot be measured with any convenient setup. This is explained in Chapter 21, which makes the case for not trying to measure the level of a scrambled signal. Rather, to measure the level, turn off the scrambling, and make the level measurement normally. Then turn the scrambling back on.

Measuring the Picture–Sound Ratio with a Spectrum Analyzer

The ratio of the picture and sound carrier amplitudes must also be set. This may be done with a spectrum analyzer set to measure picture carrier level.

Normally, in cable work, the picture–sound ratio is set to 15 dB. The off-air ratio is different, but cable uses a lower sound carrier to minimize adjacent channel interference.

The question often is asked about how to set the picture–sound carrier ratio when the signal is scrambled and pulses are used on the sound carrier. The pulses make the sound carrier level go high for a few microseconds to communicate information to the descramblers. Again, the proper way to set level is to turn off the scrambling, set the picture–sound ratio, then turn the scrambler back on. The scrambling manufacturer must ensure that the correct levels are maintained when the scrambling is turned on. Generally, the signal levels are set such that the picture–sound ratio is maintained at about 15 dB between pulses and goes to 9 or 10 dB at the peak of the pulses.

9.5.4 Measuring Digital Signals

Unique measurements to be made on digital signals are covered in Chapter 4.

9.6 Summary

This chapter has dealt with a number of somewhat disconnected topics generally related to headend operation. Three general topics were covered: the standard band plan for cable television in the United States, management of outgoing optical and electrical signals in the headend, and monitoring of the quality of signals leaving the headend.

The television receiver and cable television industries have agreed on a band plan that defines the frequency versus channel number for television channels. The plan has been defined for three sets of frequency allocations: the standard plan, used in most instances today, and the HRC and IRC plans, chosen to reduce the visibility of distortion products. In the VHF band, the standard plan is identical to the off-air FCC channel assignments. In the old midband and superband regions, the plan follows traditional assignments with numerical channel designations. In the UHF band, the cable assignments are 2 MHz below the FCC plan because the FCC plan does not move from channel 13 (210–216 MHz) to channel 14 (470–476 MHz) in 6-MHz increments.

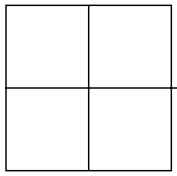
Signals in the headend must be managed in such a way that it is possible to carry different signals in different nodes but on the same frequency. This means that isolation must be maintained between unique signals bound to or from different nodes. On the other hand, certain signals are common to all nodes, and these have to be handled in common.

It is important to handle headend signals so that quality is maintained. Because the quality of signals exiting headends must be monitored, we have presented some fundamental ways to do so. Books have been written to cover the topic in detail.

We now move out of the headend and into the distribution plant. The next several chapters deal with the technology of distributing signals outside the headend.

Endnotes

1. http://global.ihs.com
2. J. O. Farmer, The Joint EIA/NCTA Band Plan for Cable Television, *IEEE Transactions on Consumer Electronics*, August 1994, p. 503ff.
3. Code of Federal Regulations, 47CFR sec. 68.610. U.S. Government publication, available at Government Printing Offices in major cities.
4. AT&T Corp., *LGX Broadband Fiber Management System Application Guide*, 1995.
5. IEEE Std 152–1991, *IEEE Standard for Audio Program Level Measurement* is one of many suggestions for measuring audio loudness and peak deviation. It includes an extensive bibliography on measuring audio, a controversial topic. IEEE, Piscataway, NJ.
6. J. O. Farmer, Setting and Maintaining Modulation Levels. *CED*, March 1987, reprinted in *BCT/E Certification Program Reference Bibliography Reprint Manual*, available from the Society of Cable Telecommunications Engineers, Exton, PA.



Chapter 10
Coaxial RF Technology

10.1 Introduction

Cable television's technical roots are the distribution of analog television signals to the antenna terminals of customers' receivers. The least expensive way to accomplish that was to avoid the need for in-home equipment by carrying each video stream on a different, standard television channel so that subscribers could use their existing TV sets to select and view programs. Although cable systems have evolved since then into sophisticated multiservice networks using combinations of linear fiber optics and coaxial cable for transmission, the essential characteristics have remained unchanged. In particular, the last portion of the network is still a linear, broadband coaxial network that simultaneously carries many modulated RF signals, each occupying a different band within the spectrum — frequency division multiplexing (FDM) — as discussed in several previous chapters. In modern systems, many of the modulating signals may be digital rather than analog; however, the network must still be linear to avoid generation of unwanted additional signals due to mixing among the desired carriers. This is discussed in detail in Section 10.3.3.

This chapter will treat coaxial network technology in detail, including cable, amplifiers, passive components, and powering systems. Basic linear network concepts will be introduced that will also apply to the linear fiber-optic links whose unique characteristics are discussed in Chapter 12 and to the microwave links covered in Chapter 14. Chapter 11 will discuss coaxial design practices and cascaded performance.

10.2 Coaxial Cable

Coaxial cable is not the only option for transmitting broadband RF signals. Indeed, many early systems were built using open, parallel-wire balanced transmission lines, and a few even used an ingenious single-wire cable

known as G-line, which had only a center conductor and dielectric. Coaxial cable, however, offers the advantages of a high degree of shielding, coupled with relatively low cost and moderate signal loss. For that reason, most modern RF systems use coaxial cable as at least one of the physical media through which signals are transported.

10.2.1 Definition

Coaxial cable is constructed with a center conductor surrounded by a dielectric of circular cross section and by an outer conductor (shield), also of circular cross section. Signals within the normal operating bandwidth of coaxial cable have a field configuration known as transverse electric and magnetic (TEM). In the TEM mode, the electric field lines go radially between the center and outer conductor and are of uniform strength around a cross section of the cable, whereas the magnetic field lines are circular and perpendicular to the length of the cable (see Figure 10.1). In a cable with a continuous, perfectly conducting shield, no electric or magnetic fields extend beyond the outer conductor so that no signals can leak out of or into the cable.

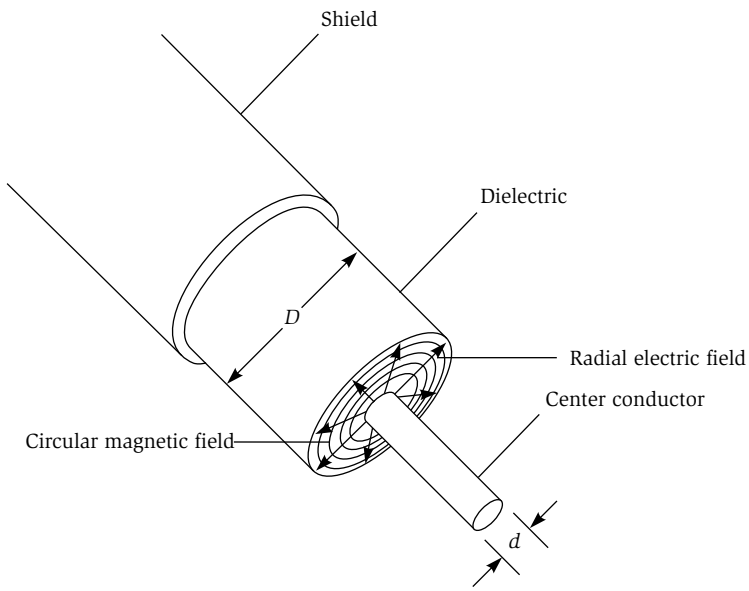


Figure 10.1 Coaxial cable basics.

10.2.2 Characteristic Impedance

Coaxial cables have a property known as surge impedance or, more commonly, characteristic impedance, which is related to the capacitance and inductance, per unit length, of the cable. The characteristic impedance is most easily thought of in terms of the effect on signals being transported: if a cable is connected to an ideal pure resistor whose value is equal to its characteristic impedance, a signal transmitted toward the resistor will be entirely absorbed by the resistor and converted to heat. In other words, no energy will be reflected up the cable.

The characteristic impedance, Z_0 (in ohms), is a function of the relative diameters of the center conductor and the inner surface of the outer conductor and of the dielectric constant of the dielectric:

$$Z_0 = \frac{138}{\sqrt{\varepsilon}} \log\left(\frac{D}{d}\right) \tag{10.1}$$

where

D = the inner diameter of the shield
d = the outer diameter of the center conductor
ε = the dielectric constant

10.2.3 Attenuation as a Function of Frequency

Signal loss (attenuation) through coaxial cable can occur through any of four principal means:

- Radiation out of the cable due to imperfect shielding
- Resistive losses in the cable conductors
- Signal absorption in the dielectric of the cable
- Signal reflection due to mismatches between the cable and terminations or along the cable due to nonuniform impedance

Even when cables have perfect shields, exact impedance matches, and uniform construction, imperfect dielectrics and resistive conductors will cause loss. The general equation for this residual cable loss is

$$\alpha = 4.344\left(\frac{R}{Z_0}\right) + 2.774F_p\sqrt{\varepsilon}\, f \tag{10.2}$$

where

α = the attenuation of the cable in dB/100 feet*
R = the effective ohmic resistance of the sum of the center and outer conductors per 100 feet of cable length at f
F_p = the power factor of the dielectric used
f = the frequency in MHz

The currents in the conductors are proportional to the strength of the magnetic fields at the conductor surface. If the conductors had no resistance, the RF currents would travel only on the surface. In real conductors, however, the current extends into the conductor, decreasing exponentially with depth. This property is known as *skin effect*, and the distance from the surface to where the current has decreased to $1/e$ (36.8%) of the surface amount is known as the *skin depth* at that frequency. It is related to frequency by the formula

$$\delta = \frac{1}{\sqrt{f}} 2.60 \sqrt{\frac{\rho}{\rho_C}} \tag{10.3}$$

where

δ = the skin depth in mils (thousandths of an inch)
ρ/ρ_C = the resistivity of the conductor relative to copper
f = the frequency in MHz[1]

The formula is valid only for nonferrous materials such as aluminum or copper.

Over a typical downstream frequency range of 54–750 MHz, the skin depth in copper will vary from 0.00035 to 0.00009 inch and will increase to 0.0012 inch at 5 MHz (the lower end of the upstream frequency range).

The effective resistance of the conductor is the same as if it were a tube of material whose thickness is equal to the skin depth with the current distributed equally throughout its volume. The resistance of this tube will be

$$R = 0.0996 \frac{\sqrt{f}}{d} \sqrt{\frac{\rho}{\rho_C}} \tag{10.4}$$

---

* In accordance with common North American practice, all attenuation values will be given in decibels per 100 feet of length. This has the advantage that signal loss becomes a linear function of length and of simple conversion to metric units of length.

where

R = the resistance per 100 feet of the conductor
ρ/ρ_C = the resistivity of the conductor relative to copper
f = the frequency in MHz
d = the conductor diameter in inches

Taking into account the resistance of both center and outer conductors, Equation (10.2) can be rewritten as

$$\alpha = \frac{0.433}{Z_0}\left(\frac{\sqrt{\rho_d/\rho_c}}{d} + \frac{\sqrt{\rho_D/\rho_c}}{D}\right)\sqrt{f} + 2.774(F_p\sqrt{\varepsilon})f \tag{10.5}$$

where

ρ_d/ρ_c = the resistivity of the center conductor material relative to copper
ρ_D/ρ_c = the resistivity of the shield material relative to copper

This has one term that increases as the square root of frequency and one that increases linearly with frequency. For most cable constructions (discussed later), the conductor loss will dominate the dielectric loss so that the overall cable attenuation will increase approximately as the square root of frequency. Figure 10.2 shows the specified variation in loss for several cables in comparison to an ideal square-root relationship (with the curves matched at 100 MHz). Note that the agreement is excellent for cables ranging from drop sizes to the largest trunk cables, indicating the relatively small contribution from dielectric losses.

A common cable construction, known as P3, utilizes a solid aluminum outer conductor, a foamed dielectric ($\sqrt{\varepsilon} = 1.15$), and a copper-coated aluminum center conductor. It is available in a number of nominal sizes (referenced to the *outer* diameter of the aluminum shield in inches). For this family of cables, the attenuation can be approximated using

$$\alpha = \left[\frac{0.036}{D_{\text{nom}}}\right]\sqrt{f} + 0.0002f \tag{10.6}$$

where

α = the attenuation in dB/100 feet
D_{nom} = the nominal shield outer diameter in inches
f = the frequency in MHz

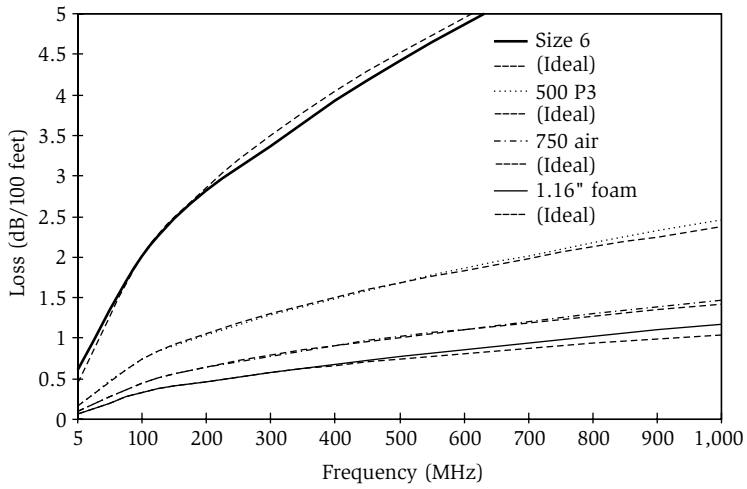


Figure 10.2 Cable loss: specified versus ideal.

Other cable types using thinner shields and/or air dielectrics will have slightly lower losses for the same outer diameter, whereas solid dielectric cables will have higher losses.

10.2.4 Attenuation as a Function of Temperature

As Equation (10.2) shows, the attenuation of a coaxial cable, in decibels, is a linear function of its conductor resistance, provided that conductor losses are much larger than dielectric losses. As we will discuss in Section 10.2.10, the most common cable construction uses an aluminum shield and copper-coated aluminum center conductor. Copper has a resistivity at 68°F of about 0.68×10^{-6} ohm-inches, whereas aluminum has a resistivity of about 1.03×10^{-6} ohm-inches, and each has a temperature coefficient of resistivity of 0.22%/°F. Since conductor resistance varies as the square root of resistivity, the resistance of both conductors, and therefore the attenuation, would be expected to change about 0.11%/°F. In fact, the attenuation of commercial coaxial cables with hybrid aluminum–copper construction is typically specified to increase by 0.1%/°F.

10.2.5 Attenuation as a Function of Characteristic Impedance

For a given outer diameter (the primary determinant of the amount of materials used and, therefore, cost and weight), different impedance cables will be optimized for different characteristics. If the dielectric loss is assumed to be of only secondary

importance and the usual construction is assumed (copper-plated center conductor with aluminum shield), then Equations (10.1) and (10.5) can be combined to find cable attenuation as a function of characteristic impedance:

$$\alpha = \frac{0.433\sqrt{f}}{D}\left[\frac{1.23 + 10^{\left(\frac{Z_0\sqrt{\varepsilon}}{138}\right)}}{Z_0}\right] \tag{10.7}$$

As Figure 10.3 shows, the loss minimum is at about 80 ohms, for air dielectric (dielectric constant = 1.0) and decreases as the dielectric constant increases. Cables with air dielectric also have lower overall loss since the center conductors are larger, and therefore have less resistance, for the same impedance. Although the choice of impedances near the minimum of the curve was not important, 75 ohms may have been chosen because it is also close to the feedpoint impedance of a half-wave dipole antenna. In order to minimize the need for repeaters, wide area distribution systems universally use 75-ohm cables. They are also used for local interconnection of baseband video signals within headends and broadcast facilities in order to minimize the differential loss across the video baseband spectrum.

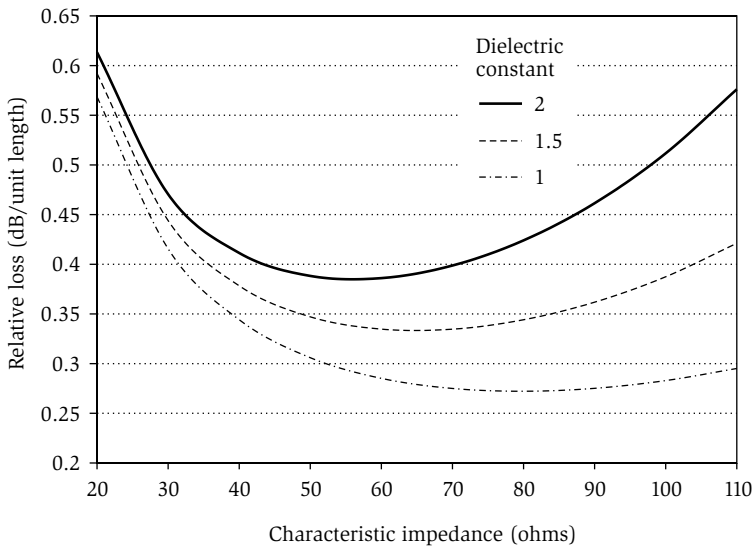


Figure 10.3 Coaxial cable loss versus characteristic impedance for various values of dielectric constant.

Although the losses are higher, 50-ohm cables are generally used by the broadcast and radio communications industries because the power-handling capability is higher, the cables are less fragile, and they are a close match to the feedpoint impedance of a vertical quarter-wave antenna.

10.2.6 Wavelength

Along the length of the cable, at any one instant in time, the electric fields vary sinusoidally in strength. At some point, the center conductor will be at its most positive with respect to the shield. Moving along the cable, the voltage will decrease to zero, then become negative, then back to zero, and finally be at its positive maximum again. The distance between the maximum points of one polarity is known as the wavelength at the frequency being transmitted and is equal to the distance the signal transverses during one period of its frequency.

In free space, signals travel at the rate of 3×10^8 meters per second or, in more convenient terms, about 984 feet per microsecond. Signals in cable, however, travel more slowly due to the higher dielectric constant of the dielectric material. The ratio between velocity in cable and the free-space velocity is the *relative propagation velocity* V_P and varies from about 0.85 to 0.95 for common cable types. V_P is related to the dielectric constant by

$$V_P = \frac{1}{\sqrt{\varepsilon}} \tag{10.8}$$

The wavelength of a signal in a coaxial cable is thus

$$\lambda(\text{feet}) = \frac{984 V_p}{f(\text{MHz})} = \frac{984}{f\sqrt{\varepsilon}} \tag{10.9}$$

10.2.7 Theoretical Size Limitation

If the size of a coaxial cable is sufficiently large, it will support modes other than TEM. In particular, if the mean diameter exceeds one wavelength in the cable, it will support a mode in which the electric and magnetic fields are not uniform around the cable. This *cutoff wavelength* can be expressed as[2]

$$\lambda_C = \pi(D + d) \tag{10.10}$$

where

D = the inner diameter of the shield
d = the outer diameter of the center conductor

Non-TEM mode generation is a problem since higher-order modes may not propagate at the same velocity and because coaxial components may not react the same to the higher-order mode. Standard practice is to limit cable sizes to those that do not support modes higher than TEM at the maximum operating frequency. The maximum size cable that will support a given frequency in TEM mode only can be found by substituting Equation (10.9) for λ in Equation (10.10), then solving Equation (10.1) for d, substituting in Equation (10.10), and solving the resultant expression for D:

$$D_{\max} = \frac{984}{\pi f \sqrt{\varepsilon}\left[1 + 10^{-\left(\frac{Z_0\sqrt{\varepsilon}}{138}\right)}\right]} \tag{10.11}$$

Alternatively, we can solve for the maximum frequency that can be transmitted in only TEM mode through a given cable:

$$f_{\max} = \frac{984}{\pi D \sqrt{\varepsilon}\left[1 + 10^{-\left(\frac{Z_0\sqrt{\varepsilon}}{138}\right)}\right]} \tag{10.12}$$

For example, 75-ohm cables with foamed dielectrics (ε about 1.3) will support 1-GHz signals at up to 0.222-foot shield inner diameter, or about 2.6 inches. Since this is more than twice the largest cable used by network operators today, higher-order mode suppression is not an issue with current operating bandwidths.

10.2.8 Precision of Match: Structural Return Loss

One measure of cable quality is how close it adheres to its nominal impedance. This has two aspects: the precision with which its *average* characteristic impedance matches the ideal value, and the *variations* of impedance along the length of a cable. As an alternative to attempting to measure the physical structure of the cable along its length, the most frequent measure of quality used is the percentage of incident power from a source that is reflected at the input of a cable being tested, when the cable's output is connected to a precision termination. If the source is calibrated using a precision, resistive terminator, then the result of the measurement is return loss. For cables, however, the variation of the reflection is often of more interest than the absolute match. In order to judge the variation, a variable bridge is used to match, as precisely as possible, the average surge impedance of the cable being tested. Then the reflection is measured over the full frequency range of desired operation, and the result is known as structural return loss (SRL). Its numeric value is the smallest measured ratio of incident to reflected power as the frequency is varied, expressed in decibels. Recommended worst-case SRL is 26–30 dB for trunk and feeder cable, whereas

20 dB is considered adequate for drop cable. The absolute impedance should be within $\pm 2-3$ ohms for trunk and distribution cables and ± 5 ohms for drop cables.[3]

SRL is a compromise measurement, as the reflected power is a measure, at each frequency, of the composite effect of discontinuities at all distances. Because of the attenuation of the cable, signals traveling toward more distant discontinuities will be reduced in amplitude, and the reflections from those discontinuities will be further attenuated so that imperfections farther from the end of the cable being measured will have less effect than those closer. At higher frequencies, this effect will be increased. On the other hand, most cable imperfections cause a greater reflection at higher frequencies.

Where multiple reflections are spaced along a transmission line, the effects can add, cancel, or anything in between, depending on the length of line between them and the frequency. (Chapter 15 includes a more complete discussion of multiple reflections in coaxial transmission systems.) Absent line losses, small, identical discontinuities will approximately cancel when spaced odd multiples of a quarter wavelength, whereas they will reinforce when spaced multiples of one-half wavelength. This is especially important when considering multiple, equally spaced discontinuities arising, for instance, from an imperfection in a roller over which a cable passes during the manufacturing process. Each revolution of the roller produces an identical fault, spaced along the cable by a distance equal to the circumference of the roller. Testing such a cable over the full frequency range will show maximums and minimums of return loss as the effects of the discontinuities interact. This is why it is important to test cable over its full expected frequency range with sufficient frequency resolution to detect sharply resonant discontinuities.

Though structural return loss measurements accurately assess the composite reflection from a cable, it is a poor method of determining the location of major defects. If a cable includes only a single major defect, its location can be estimated from the change in test frequency between successive maxima or minima of return loss; however, in most cases, cable loss plus a multiplicity of reflections will prevent accurate determination.

10.2.9 Precision of Match: Time Domain Reflectometry (TDR)

An alternative measurement method, also a compromise, is used to more accurately determine discontinuity location but averages the magnitude of the discontinuity as a function of frequency. In this test, a pulse is transmitted down the cable, and the distance is determined from the time required for the energy reflected from a discontinuity to return to the transmitter. The magnitude of the discontinuity is related to the amplitude of the returned signal. Secondary information about the frequency "signature" of the fault can be determined by comparing the shape of the returning pulse with the original.

Although TDR testing of short, low-loss cables is an accurate process, there are many compromises in testing of the relatively long cables used in typical cable system construction. With very short cables, the classic pulse waveform is a very fast rise time step function. A Fourier transform of this voltage waveform reveals that its frequency content is inversely proportional to frequency, and therefore, the power spectrum falls inversely as the square of frequency (see Figure 10.4(a)). When such a pulse is used to test long cables, it suffers from two problems: (1) most of the energy is concentrated at low frequencies, whereas most discontinuities primarily affect higher frequencies, and (2) the relatively higher cable loss at high frequencies further masks the reflections.

For that reason, long cables are most frequently tested using a narrow "impulse" (see Figure 10.4(b)). This waveform has a spectrum that extends more uniformly to a frequency equal to approximately $1/P_W$, where P_W is the pulse width (the exact shape of the curve in the frequency domain depends on the shape of the voltage pulse). The trade-off is that the distance resolution is limited to approximately $cP_W V_P$, or about 1 foot per nanosecond of pulse width for typical foam dielectric cables. By increasing P_W, it is possible to increase the total energy (and therefore sensitivity) at the expense of frequency and distance

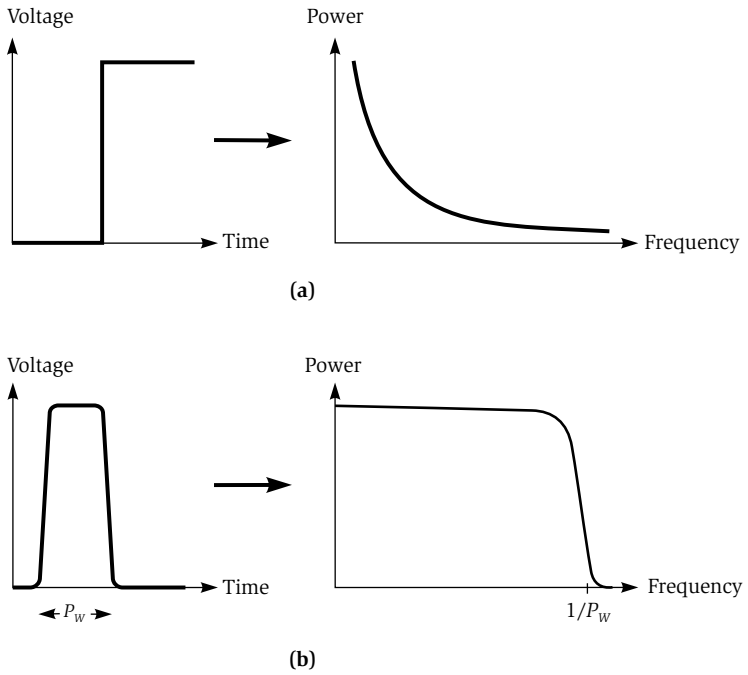


Figure 10.4 Frequency spectra of step and pulse TDR waveforms. (a) Step function. (b) Pulse.

resolution. As with step function pulses, impulse function testing suffers in accuracy because of the variation of cable loss with frequency. Because of the compromises, TDR testing of CATV-type cables is primarily limited to field operations such as measuring the cable remaining on a reel or locating an underground cut so that the cable can be repaired.

An instrument that is able to simultaneously resolve both frequency and time (distance) domain information is a network analyzer with gated TDR and Fourier transform capability. This instrument measures reflection coefficient, but "gates" the measurement so as to measure only reflections from a narrow range of distances at a time. A Fourier transform is then used to convert the waveform of the reflected signal into its frequency domain, resulting in a reasonably accurate measurement of the reflection coefficient of an individual discontinuity, even in the presence of other discontinuities at other distances. It is still, of course, affected by the cable loss and loss variation with frequency. As of the writing of this book, this technology has not been extensively used in field work in cable systems.

10.2.10 Practical Factors in Cable Selection

From an electrical standpoint, the ideal cable would have conductors with the highest possible conductivity, the largest diameter consistent with higher-order mode suppression, and a vacuum separating the conductors. Practical cables, however, are a compromise among several characteristics:

- Manufacturing cost (which will vary somewhere between being a linear function of diameter and as the square of the diameter)
- Weight (which affects the cost of the supporting structure for overhead lines)
- Diameter (which affects required conduit size for underground lines, visibility of overhead lines, and cost of connectors)
- Loss (which limits the distance between repeater amplifiers)
- Mechanical expansion coefficient (which affects the amount of sag that must be used with overhead lines)
- Connector cost (related to cable size)
- Shielding (although solid outer conductors provide nearly perfect isolation, braided shields may provide adequate shielding for some applications)
- Environmental protection (mechanical protection layers, compounds used to retard water ingress, integral support members, and so on)
- Handling characteristics (bending radius, reforming ability, crush resistance, pulling force resistance, and so on)
- Electrical resistance and current-carrying capacity at low frequencies

In oversimplified terms, the selection of optimum cable size for a distribution network design consists of comparing the cost of cable (and associated components) per decibel of loss with the cost per decibel of amplification to overcome that loss and choosing the least expensive combination. Of course, practical system design includes many additional factors that may affect the ultimate decision.

Trunk and Feeder Cable

Design optimization has resulted in several typical cable designs that are widely used. The most common distribution cable design in most of the world utilizes a solid aluminum shield, foamed polyethylene dielectric, and a copper-clad aluminum center conductor. The use of aluminum offers the best combination of cost, handling characteristics, weight, and strength, whereas the foamed dielectric offers a good compromise among loss, moisture ingress protection, and mechanical strength for the center conductor. Because of its rigidity, compared with cables that utilize braided wire shields, trunk and feeder cables using this basic structure are sometimes known as *hard-line* cable. The use of an aluminum core for the center conductor provides a match for the expansion coefficient of the shield, whereas the copper cladding provides a lower resistance to RF. Since the conducting area of the center conductor is much smaller than the shield, the total resistive losses are dominated by it so that copper cladding the center conductor is cost efficient though coating the inner side of the shield would not be.

The need to provide low resistance to the multiplexed power, as well as low RF loss, has led to the use of cables with outer shield diameters ranging from 0.412 to 1.125 inches, with larger cables being used where distance of coverage is paramount (trunk or express feeders), and smaller cables being used for shorter distances and cases where a substantial amount of the loss is due to tap losses rather than cable losses (tapped feeder cables). This family of cables is generically labeled by the outer shield diameter in thousands of an inch; for example, "500 cable" refers to a hard-line cable with an outer shield diameter of 0.500 inch. Common diameters include 0.412, 0.500, 0.625, 0.750, 0.825, and 1.000 inch though others are also used.

Versions of these same cables are available with solid copper center conductors, for use where 60-Hz power losses must be reduced. Protection options include combinations of none (bare aluminum), polyethylene jackets of various designs and degrees of sun and fire resistance, flooding compounds (a substance inserted between the shield and jacket to retard the ingress of water), and additional steel shields for crush resistance. Since the added environmental protection layers can considerably increase the cost and weight, builders generally utilize several designs in a single system, optimized by application.

A variation of this design utilizes air dielectric with periodic center conductor supports. This design has slightly smaller diameter for the same loss per foot

because of the lower dielectric loss and lower ratio of shield to center conductor diameter. Special design features are required to prevent water, which may enter the cable owing to a fault, from traveling down the center conductor/shield space. Other features are required to maintain the required concentricity and crush or bending resistance.

Though hybrid aluminum–copper cables are the most common, some network operators prefer solid copper for center conductors and, sometimes, shields.

Drop Cable

In a typical cable system, well over half the total cable footage is taken up by drop cables. For that reason, cost is an important parameter for that application, as well as weight and appearance. In North America, four common sizes of drop cable have evolved, as shown in Table 10.1. The approximate loss for each size, as a function of frequency, is shown in Figure 10.5.[4]

Table 10.1 Common drop cable designations and sizes

Size Designation	*59*	*6*	*7*	*11*
Approximate diameter over jacket in inches	0.24	0.27	0.34	0.40

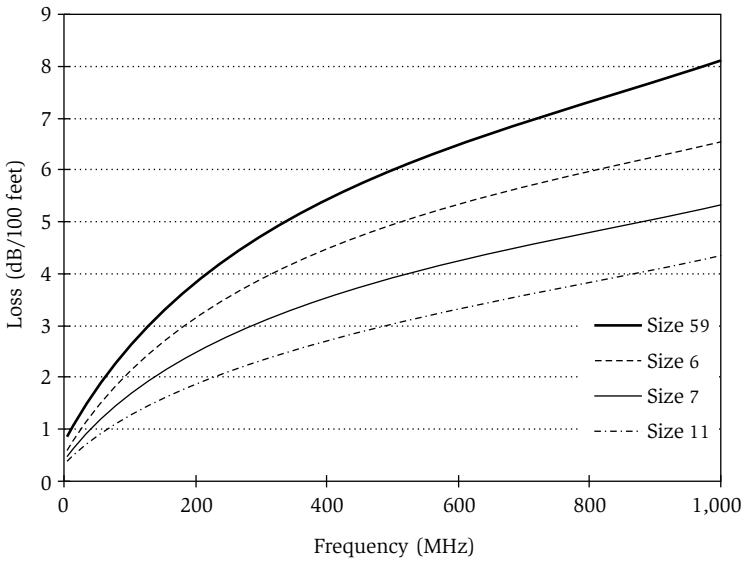


Figure 10.5 Drop cable loss versus frequency.

Drop cables are constructed using a copper-plated steel center conductor, foamed dielectric, a shield made up of at least one layer of an aluminum foil and one layer of an aluminum braid, and a plastic jacket. Additional layers of foil and braid are available as options for greater shielding effectiveness. Although size 59 cable was the historical choice of the cable television industry, size 6 cable is most frequently used in modern systems, due to the use of increasingly higher frequencies. The larger sizes are generally reserved for occasional longer drop cables.

The jacket material for general use is usually PVC or polyethylene. Cables used for overhead drops often also include a separate steel strength member (the *messenger*) enclosed in the same plastic sheath structure. See Figure 10.6 for a composite drop cable structure that includes a messenger.

An array of drop cable variations are available for special purposes, including those with *flooding compounds* to retard water ingress in case the plastic

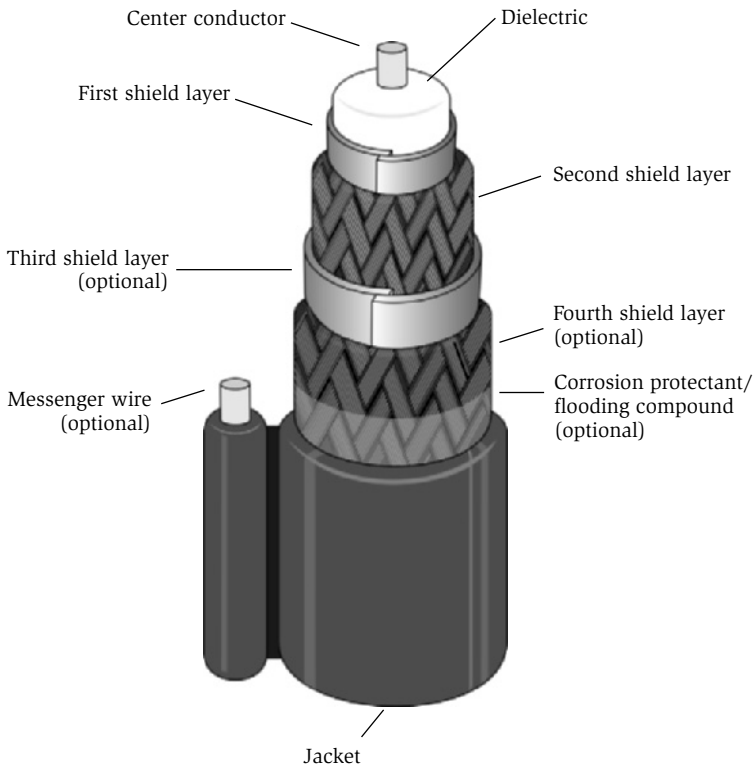


Figure 10.6 Typical drop cable construction.

coating is damaged, multiple cables within a common jacket, those with varying degrees of fire resistance for use inside buildings, silver-plated center conductors, copper braids, solid dielectric cables, and so on.

As with distribution cables, North American practice has been adopted by some countries, whereas others (for instance, Germany) have chosen to use semirigid drop cables with solid copper shields and center conductors for their underground drops.

10.2.11 Shielding Effectiveness in Drop Cables

Although distribution cables constructed with solid aluminum tube shields offer nearly perfect signal isolation from external fields, drop cables depend on the overlap in the aluminum wrap and the interaction between the aluminum braid and wrap for shielding integrity. It is a compromise that is justified on the basis of cost and cable-handling characteristics. The cable industry has long struggled to find nonambiguous testing methodology for evaluating shielding integrity. Although these are beyond the scope of this book, you may wish to refer to industry standard test procedures developed by the Society of Cable Telecommunications Engineers (SCTE) for this purpose.[5]

10.3 Amplifiers

In order to overcome the loss of the distribution cables and to provide power to drive end terminals, signal boosting amplifiers are utilized throughout the coaxial distribution system. These must be designed to add as little signal degradation as possible, particularly noise and distortion, consistent with providing the required gain and total power output.

10.3.1 Broadband Random Noise

Because of the random motion of electrons in conductors, all electronic systems generate an irreducible amount of noise power. This noise is a function of the absolute temperature and the bandwidth in which the noise is measured. As discussed in Chapter 7, the minimum thermal noise power (the *noise floor*) can be calculated using

$$n_p = kTB \tag{10.13}$$

where

n_p = the noise power in watts*

k = Boltzmann's constant (1.374×10^{-23} joules/°K)

T = the absolute temperature in °K

B = the bandwidth of the measurement in Hz

Signal power in cable television systems is commonly expressed in decibels relative to 1 mV† across 75 ohms (dBmV), so the thermal noise power at 62°F is approximately

$$N_p(\text{dBmV}) = -125.2 + 10 \log B \tag{10.14}$$

Finally, although the allocated channel bandwidth for analog NTSC signals is 6 MHz, the effective noise bandwidth of receivers is less. For a variety of reasons, a 4-MHz bandwidth is used as the standard for noise measurements of analog NTSC signals and has been codified in the FCC's rules.[6] At that bandwidth, Equation (10.14) becomes

$$N_p = -59.2 \text{ dBmV} \tag{10.15}$$

This value will vary by a few tenths of a decibel across the temperature range normally of interest to cable operators.

10.3.2 Amplifier Noise

Amplifiers generate added noise at various points in their circuitry. For convenience, however, the added noise is treated as if it were coming from an independent generator and summed into the input port (though it cannot be measured there, of course). The ratio of total effective input noise power to the thermal noise floor, expressed in decibels, is known as the *noise figure* of the amplifier. Thus, an amplifier with a noise figure of F_A dB will have a total equivalent input noise power of N_A, where

$$N_A \text{ (dBmV)} = N_p \text{ (dBmV)} + F_A \text{ (dB)} \tag{10.16}$$

---

* To avoid confusion, lowercase symbols will be used in this discussion to refer to scaler quantities (for example, watts), and uppercase symbols will refer to logarithmic quantities (for example, dBmV).

† In some countries, the reference is 1 μV rather than 1 mV. Levels in dBμV are therefore numerically greater by 60 than levels in dBmV. In either case, however, the reference is power, not voltage. Finally, for the convenience of those accustomed to working in systems referenced to 1 mW (dBm), 0 dBmV = −48.75 dBm. For consistency, this book uses dBmV as the units for RF power unless otherwise specified.

And the output noise power will be the input, increased by the gain, G, of the amplifier in decibels:

$$N_{\text{out}}\ (\text{dBmV}) = N_p + F_A + G \tag{10.17}$$

Similarly, a desired (noise-free) input signal, C_i, will be amplified by the same amount so that the desired signal output level $C_{\text{out}} = C_i\ (\text{dBmV}) + G\ (\text{dB})$, and the carrier-to-noise ratio will be*

$$\begin{aligned} \text{C/N}\ (\text{dB}) &= (C_i + G) - (N_p - F_A + G) \\ &= C_i - N_p - F_A \end{aligned} \tag{10.18}$$

When evaluated relative to the 4-MHz bandwidth commonly used for NTSC C/N measurements, this becomes

$$\text{C/N}\ (\text{dB}) = C_i\ (\text{dBmV}) + 59.2 - F(\text{dB}) \tag{10.19}$$

Note that, in an FDM system, C/N is evaluated on each signal independently, with the level of individual carriers compared with the noise in the bandwidth that affects that signal. As will be seen, the C/N may be considerably different for different carriers in the spectrum. (The design of systems with cascaded amplifiers will be treated in Chapter 11.)

10.3.3 Distortion

Amplifiers are not perfectly linear. That is, if the instantaneous output voltage is plotted against the instantaneous input voltage, the graph will not be a perfectly straight line. There are two causes of this nonlinearity: the inevitable minor small-signal nonlinearities of the semiconductor devices used and the compression that takes place as the amplifier nears its saturation voltage. Though the small variations define a distortion "floor," most of the distortion is due to saturation effects. They can take any of three related forms: even-order distortion, odd-order distortion, and cross modulation.

---

* As a reminder, the term C/N (carrier-to-noise ratio) is used in this book for RF modulated signals, whereas S/N (signal-to-noise ratio) is used for measurements of the noise level on baseband (unmodulated) signals. This is consistent with both regulatory terminology and common industry practice.

Even-Order Distortion (Composite Second Order—CSO)

A perfectly linear amplifier has a transfer function that is a constant; that is, the output voltage is a linear function of the input voltage. If the transfer function is not linear, then, in general, it can be expressed as a power series; that is,

$$e_o = Ae_i + Be_i^2 + Ce_i^3 + De_i^4 + \cdots \tag{10.20}$$

where

e_o = the output voltage
e_i = the input voltage
$A, B, C, D, \ldots$ = the gains at the various powers of the input voltage[7]

Those terms with even-numbered powers are called even-order distortion, and those with odd-numbered powers are called odd-order distortion. They act in very different ways in cable systems utilizing the most common channelization plans.

If a single sine wave signal $[e_i = E_i \sin(\omega t)]$ is transported through an amplifier with second- and fourth-order distortion, the output voltage will be

$$e_o = A[E_i \sin(\omega t)] + B[E_i \sin(\omega t)]^2 + D[E_i \sin(\omega t)]^4 \tag{10.21}$$

where

ω = the angular frequency
A = the linear component of gain
B = the amplitude of the squared nonlinear component
D = the amplitude of the fourth power component
E_i = the peak input voltage

Using standard trigonometric identities, this is equivalent to

$$e_o = \frac{BE_i^2}{2} + \frac{3DE_i^4}{8} + AE_i \sin(\omega t) - \frac{(BE_i^2 + DE_i^4)}{2}\cos(2\omega t) + \frac{DE_i^4}{8}\cos(4\omega t) \tag{10.22}$$

Analyzing this expression, we find that

- The first two terms of this expression represent a change in the average dc voltage that may correspond to a change in the power dissipation of the active device, but is not transmitted into the ac-coupled transmission system.

- The third term represents the desired output signal. Note that this amplitude does not change as the result of the even-order distortion.
- The fourth term represents an additional signal at twice the frequency of the input signal and in phase with it $\omega = \pi/2$ radians. In real amplifiers, B is substantially greater than D. Thus, for small values of E_i, the amplitude of this factor will increase as the square of the input voltage. At some input level, the fourth power term will begin to dominate, and the second harmonic level will increase more quickly. Since the level of the fundamental increases linearly with input signal level, the ratio of the fundamental to the second harmonic, at low levels, is a linear function of input level. Expressed in logarithmic terms, the "second-order ratio" degrades by 1 dB for every decibel increase in operating level so long as the amplifier is "well behaved" (meaning the fourth power terms are insignificant relative to the second power terms).
- The final term is the fourth harmonic of the fundamental. Like the fourth power term of the second harmonic, this product is generally small at normal operating levels and at all levels is at least 12 dB (one-fourth the voltage) below the second harmonic term. Note that this term is also in phase with the fundamental at $\omega = \pi/2$.

If we plot the fundamental and first two even harmonics on a time scale (Figure 10.7), we see an interesting relationship: the distortion-caused products are *antisymmetrical* with respect to the fundamental; that is, whatever voltage they have at the positive peak of the fundamental, they have exactly the opposite during the negative peak. The usual case of the combined composite signal is shown in the bold line in the figure, which shows one peak of the composite waveform flattened relative to the fundamental, whereas the other peak is sharpened.

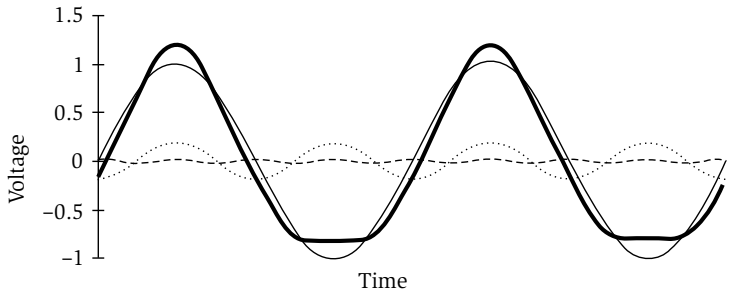


Figure 10.7 Even-order distortion.

As discussed in Chapter 9, the North American Standard channel plan (ANSI/EIA-542) places most analog television visual carriers approximately 6 MHz apart and offset approximately 1.25 MHz upwards from harmonics of 6 MHz.* Each such carrier can be expressed as

$$\sin[2\pi(6n + 1.25)t] \tag{10.23}$$

where the angle is expressed in radians and

$$n = \text{the closest harmonic of 6 MHz}$$
$$t = \text{the time in microseconds}$$

When two such signals (on different channels) are transported through an amplifier with second-order distortion (for simplicity, and because the effects are often small, the fourth-order term has been dropped), the output is

$$\begin{aligned} e_o = {} & AE_i(\sin[2\pi(6n + 1.25)t] + \sin[2\pi(6m + 1.25)t]) \\ & + BE_i^2(\sin[2\pi(6n + 1.25)t] + \sin[2\pi(6m + 1.25)t])^2 \end{aligned} \tag{10.24}$$

where the first term is the linear amplification of the input signals and

$$n \text{ and } m = \text{the multiples of 6 MHz closest to the two visual carriers}$$
$$E_i = \text{the peak voltage of either of the input signals (they are assumed to have equal amplitudes)}$$

The last term has a dc component plus the following intermodulation products:

$$-BE_i^2 \cos[2\pi(6[m + n] + 2.5)t]$$
$$+BE_i^2 \cos[2\pi(6[m - n])t]$$
$$-(BE_i^2/2) \cos[2\pi(12n + 2.5)t]$$
$$-(BE_i^2/2) \cos[2\pi(12m + 2.5)t]$$

The first product falls 1.25 MHz above the visual carrier nearest the $(m + n)$th harmonic of 6 MHz, whereas the second product falls 1.25 MHz below the visual carrier nearest the $(m - n)$th harmonic of 6 MHz.

---

* Variations from exact spacing are due to equipment frequency tolerances, broadcast station offsets of 10 or 20 kHz, and aeronautical band offsets of 12.5 or 25 kHz.

The third and fourth products will be half the voltage (−6 dB in power) and will each lie 1.25 MHz above another visual carrier (provided, in all cases, that the products fall within the occupied bandwidth of the distribution system). Note that these last products are identical in amplitude to those analyzed with a single carried signal and are, in fact, just the second harmonics of the two original signals.

The amplitude of each of these four distortion products, as with the single channel analysis, depends on the square of the input voltage so that the *ratio* of fundamental to second-order products varies linearly with input level. A similar analysis for fourth-order distortion, the next even power, will show low amplitude components at four times the original frequencies and additional components at twice the original frequencies.

When many similarly spaced carriers in the Standard channel plan are subjected to second-order distortion, each of the individual carriers and each of the carrier pairs will mix as analyzed above. This will result in clusters of products 1.25 MHz above and below each of the original carriers.* As a general rule, the number of beats falling closest to a specific carrier can be calculated using the expression

$$N_L = n - m - x + 1 \quad N_U = \text{INT}\left[\frac{x - 2m + 1}{2}\right] - \begin{cases} 0 & \text{if } x \le n \\ x - n - 1 & \text{if } x > \text{n} \end{cases} \tag{10.25}$$

where INT indicates that the expression is to be rounded down to the nearest integer and

N_L = the number of lower $(A - B)$ beats
N_U = the number of upper $(A + B)$ beats
n = harmonic number of highest carrier
m = harmonic number of lowest carrier
x = harmonic number of carrier being evaluated

Since, in the ANSI/EIA-542 Standard channel plan, the carriers are not exactly spaced, the second-order products are not frequency coherent, but rather are clustered around the nominal product frequencies.

The usual measurement technique when evaluating this type of distortion is to use a filter bandwidth of 30 kHz, which is wide enough to contain all the

* Actually, there are also some products spaced ±750 kHz from the visual carriers because two of the channels (5 and 6) are offset from the comb in the Standard channel plan.

significant products in a group and to measure the total apparent power within the filter bandwidth. The power level of the upper composite beat cluster, relative to the sync peak power of the visual carrier, is known as composite second order (CSO) and is sometimes expressed in units of dBc (meaning decibels relative to the carrier power level, which is always a negative number). Some engineers prefer the notation C/CSO, the ratio of the carrier to beats in decibels (a positive number). This book will use this notation. The absolute values are the same in either case. Although the lower CSO cluster will generally have greater magnitude than the upper cluster on lower-frequency channels, it falls at the edge of the video channel and is often ignored since it has a minor effect on video quality compared with the in-band upper cluster.

Since analog video carrier modulation results in average signal levels that vary with program content, equipment is specified and evaluated when loaded with unmodulated carriers of a defined level in order to achieve repeatable measurements. When loaded with video carriers whose peak level is the same as the unmodulated levels used for the test, the average carrier power will be about 6 dB lower, resulting in an improvement of about 6 dB in C/CSO.

When the IRC channel plan is used, the visual carrier frequencies are derived from a common source so that all channels are phase coherent and offset from harmonics of 6 MHz by exactly the same amount. The result is that the second-order beats are also coherent.

When the HRC channel plan is used, the visual carrier frequencies are derived from a common source and *not* offset. Under those circumstances, the upper and lower second-order beats fall exactly on the visual carrier frequencies, and the difference frequency beats cannot be ignored.

The amplitudes of the individual products and the number in each cluster are a function of the nonlinearity of the amplifier, the levels of the carriers generating the beat, and the number of test signals used. Although the preceding analysis assumed a uniform carrier level, standard operating practice is to vary the carrier power in a controlled way as a function of frequency. This will result in varying individual beat levels.

When a comb of signals is subject to second-order distortion, the greatest number of products will fall on the highest and lowest channels. Figure 10.8, as an example, gives the total number of upper and lower 1.25-MHz second-order products in the Standard channel plan (in which channels 5 and 6 are offset by 2 MHz, and the 15th to 17th harmonics are not used).

The visual effect of the upper-side second-order products on analog NTSC television pictures is similar to that of a discrete interfering carrier — that is, closely spaced diagonal lines as discussed in Chapter 2. As discussed earlier, the lower CSO beats have relatively little effect on visual quality.

Many simple broadband RF amplifiers are single-ended class A devices, meaning that their output stages consist of a single active device that is biased to the approximate middle of its dynamic range. Such a device typically exhibits an asymmetrical compression behavior so that second-order distortion products

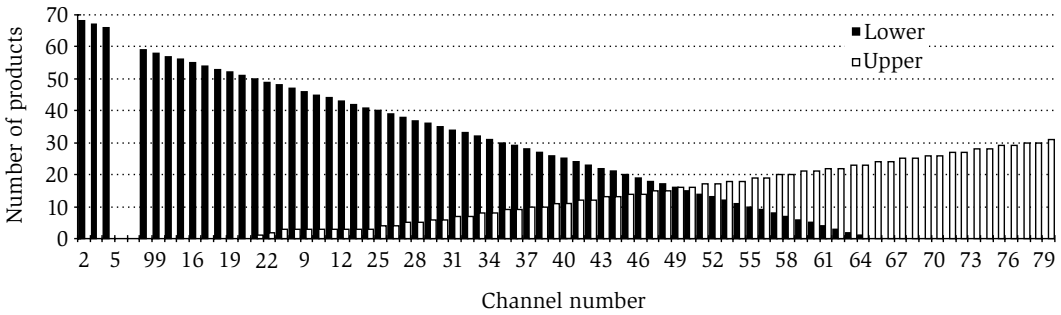


Figure 10.8 Number of 1.25-MHz CSO beats: 550-MHz Standard channel plan.

dominate. In recognition of this problem, the North American VHF off-air frequency plan designated channel boundaries such that neither second harmonic nor second-order beats from any combination of channels affect any other channels even if the drive levels to television receivers are high enough to result in compression. This was discussed in detail in previous chapters.

Early cable television amplifiers used single-ended designs for the same reasons. As the industry began using non–off-air channels, however, CSO quickly became the primary performance-limiting parameter.

Odd-Order Distortion (Composite Triple Beat—CTB)

In an attempt to reduce this problem, all modern cable amplifiers are "push–pull," meaning that their output stages contain two devices in a balanced circuit that assures symmetrical, or nearly symmetrical, compression. As will be seen, this will result in primarily odd-order distortion products.

If a single sine wave is transmitted through a circuit that has only third-order distortion, the output can be described by the following equation (derived from Equation 10.20):

$$e_o = Ae_i + Ce_i^3 \tag{10.26}$$

where A and C are the coefficients of the linear and cubed products, respectively, and

$$e_i = E_i \sin(\omega t)$$

Using a standard trigonometric identity, this reduces to

$$e_o = \left[AE_i + \frac{3CE_i^3}{4}\right][\sin(\omega t)] - \left[\frac{CE_i^3}{4}\right][\sin(3\omega t)] \tag{10.27}$$

C may be either positive or negative. In the case where the nonlinearity is due to compression, it is negative so that the third-order distortion results in a reduction in the amplitude of the fundamental (unlike second-order distortion that does not affect the amplitude of the fundamental) plus the addition of a product at three times the original frequency whose peaks are in opposition to those of the fundamental. The level of the third-order products rises as the cube of the input voltage (e.g., the rise is 3 dB for every 1-dB rise in input level) so that the *ratio* of fundamental to third harmonic at the output decreases by 2 dB for every 1-dB rise in the input level *so long as the ratio between them is high*. When the reduction in the amplitude of the fundamental becomes material, the ratio will decrease more quickly.

Figure 10.9 shows visually the fundamental, third harmonic, and the sum of the two. As can be seen, third-order distortion is equivalent to a reduced amplitude, symmetrically flattened waveform. If higher odd multiples of the original frequency are also plotted on the graph, it will be seen that all such harmonics will have exactly the same effect on both peaks of the original waveform.

If multiple signals are transmitted through a system exhibiting this type of distortion, the input can be described by the following general formula:

$$e_i = E_1 \sin[\omega_1 t] + E_2 \sin[\omega_2 t]) + \cdots + E_n \sin[\omega_n t] \tag{10.28}$$

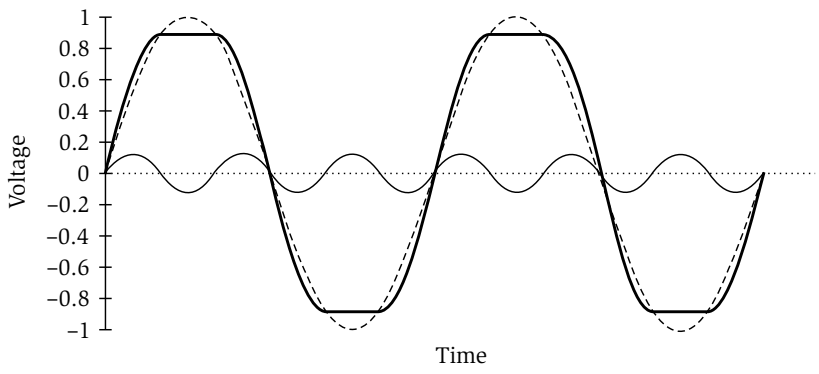


Figure 10.9 Effect of third-order distortion.

where

$\omega_{1,2,\ldots,n}$ = the angular frequencies of the input signals
$E_{1,2,\ldots,n}$ = the levels of each subcarrier (peak voltage)

If we insert this definition of e_i into Equation (10.26), it can be shown that the output spectrum will have signals at the fundamental frequencies of the input signals and all combinations of $3\omega_x$, $2\omega_x \pm \omega_y$, and $\omega_x \pm \omega_y \pm \omega_z$, where ω_x, ω_y, and ω_z are any of the input frequencies. In particular, single frequency terms, two frequency terms, and three frequency terms will be present:

- Single frequency terms:

$$\left[AE_x + \frac{3CE_x^3}{4}\right][\sin(\omega_x t)] - \left[\frac{CE_x^3}{4}\right][\sin(3\omega_x t)] \tag{10.29}$$

  Just as with a single sine wave, this indicates that the output contains, in addition to the linearly amplified input signals, two terms proportional to the cube of the input voltage level of each of the carriers, one at the fundamental carrier frequency and one at three times the carrier frequency. When the input consists of carriers corresponding to the EIA Standard or IRC channel schemes (offset 1.25 MHz from harmonics of 6 MHz), the third harmonic term will fall 3.75 MHz above a harmonic of the root frequency (2.5 MHz above a carrier if one is present). With carriers that fall exactly on harmonics of a common root (HRC), the distortion products will also fall on harmonics of the root. The number of single frequency terms will be the same as the number of FDM carriers.

- Two frequency terms:

$$\left[\frac{3CE_x^2E_y}{2}\right]\sin(\omega_y t) - \left[\frac{3CE_x^2E_y}{4}\right]\sin([2\omega_x + \omega_y]t) + \left[\frac{3CE_x^2E_y}{4}\right]\sin([2\omega_x - \omega_y]t) \tag{10.30}$$

  Each unique ordered pair of input signals mixes to produce a product at the fundamental of one of the input signals plus products at $2\omega_x \pm \omega_y$. The amplitude of the fundamental is twice that produced by the cubed term (assuming equal amplitude input signals). The voltage amplitudes of the other two terms are three times that of the cubed terms (+9.5 dB). When the input consists of carriers corresponding to the EIA Standard or IRC channel schemes, the $2\omega_x + \omega_y$ term will fall 2.5 MHz above a carrier assignment, while the $2\omega_x - \omega_y$ term will fall on a carrier assignment. With an HRC

channel assignment scheme, all products will fall directly on channel frequencies.

When the input signals consist of an FDM spectrum of q input carriers, the number of ordered pairs is $q(q-1)$. Since each ordered pair results in the generation of products at two new frequencies, the number of new products will be $2q(q-1)$. Thus, when the number of carriers is large, the total number of products will increase approximately as the square of the channel count.

- Three frequency terms:

$$\left[\frac{3E_xE_yE_z}{2}\right]\left[\sin([\omega_x+\omega_y-\omega_z]t)+\sin([\omega_x-\omega_y+\omega_z]t)\right.$$
$$\left.-\sin([\omega_x+\omega_y+\omega_z]t)-\sin([\omega_x-\omega_y-\omega_z]t)\right] \tag{10.31}$$

Each unique combination of three input signals mixes to produce terms at $\omega_x \pm \omega_y \pm \omega_z$. When the input signal amplitudes are equal, the voltage amplitudes of these terms are six times that of the cubed terms (+15.6 dB). When the input consists of carriers corresponding to the EIA Standard or IRC channel schemes, the $\omega_x+\omega_y-\omega_z$ and $\omega_x-\omega_y+\omega_z$ terms will fall on a carrier frequency, the $\omega_x+\omega_y+\omega_z$ term will fall 2.5 MHz above a carrier frequency, and the $\omega_x-\omega_y-\omega_z$ term will fall 2.5 MHz below a carrier frequency. Note that, though $\omega_x \pm \omega_y \pm \omega_z$ is sometimes negative, the only effect is the polarity of the magnitude since $\sin(-x) = -\sin(x)$.

When the input signals consist of an FDM spectrum of q input carriers, the number of unique three-signal combinations is $q(q-1)(q-2)/6$. Since each of these combinations results in the generation of four new products, the total number of new third-order products caused by three-signal mixing will be $4q(q-1)(q-2)/6$. Thus, when the number of carriers is large, the total number of products will increase approximately as the cube of the channel count.

In summary, when an FDM spectrum of q signals is transmitted through an amplifier exhibiting third-order distortion, the original signals will be amplified, and a number of new products will be generated:

- Some that affect the output amplitude of each original carrier.
- q new signals, one at the third harmonic of each of the original signals. With the ANSI/EIA-542 Standard or IRC frequency plan, these fall 2.5 MHz above visual carrier frequencies.
- $2q(q-1)$ new signals at frequencies $2\omega_x \pm \omega_y$, where ω_x and ω_y are any two input carrier frequencies. Each of these signals is about 9.5 dB higher in level than the third harmonic frequencies. With the Standard or IRC frequency

plan, half of these fall 2.5 MHz above, and half nominally on, visual carrier frequencies.

- $4q(q-1)(q-2)/6$ new signals at frequencies $\omega_x \pm \omega_y \pm \omega_z$, where ω_x, ω_y, and ω_z are any three different carrier frequencies. Each of these products will be about 15.6 dB higher in level than the third harmonic frequencies. With the Standard or IRC frequency plan, half of these will fall nominally on visual carrier frequencies, whereas one-quarter will be 2.5 MHz below and one-quarter will be 2.5 MHz above visual carrier frequencies.

The level of each of these products increases by 3 dB for every 1-dB increase in the levels of the input signals so that the ratio between the desired output signals and third-order products decreases by 2 dB.

Since the three-signal products ("triple beats") are the greatest in both number and magnitude, they quickly dominate other third-order effects. Since half of them fall nominally on visual carrier frequencies, those product groups are always the largest. Thus, composite triple beat (CTB) distortion is defined as the total power in a 30-kHz bandwidth centered on the visual carrier frequency without the visual signal present, relative to the peak level of the visual carrier in dBc (decibels relative to a reference carrier level). As with CSO, it can also be expressed as the ratio of visual carrier level to apparent beat power, or C/CTB, in decibels.

For a contiguous set of equally spaced carriers, the number of third-order beats falling in any given channel is approximately[8]

$$N_{CTB} = \frac{(a-1)^2}{4} + \frac{(a-b)(b-1)}{2} - \frac{a}{2} \tag{10.32}$$

where

N_{CTB} = the number of third-order beat products whose frequencies are nominally the same as the carrier being evaluated

a = the total number of contiguous, equally spaced carriers

b = the index number of the carrier being evaluated, i.e., the bth carrier from the bottom or top

For the center channel, which receives the greatest number of beats, this reduces to

$$N_{CTB} = \frac{3a^2 - 10a + 2}{8} \approx \frac{3a^2}{8} \qquad \text{for } n \gg 1 \tag{10.33}$$

At the end channels, which receive the fewest number of beats, it becomes

$$N_{CTB} = \frac{a^2 - 4a + 1}{4} \approx \frac{a^2}{4} \qquad \text{for } n \gg 1 \tag{10.34}$$

Actual cable systems using the EIA Standard channel plan, however, do not conform exactly to this model, with channels 4 and 5 being offset and some or all of the channels falling in the FM broadcast plan seldom used for video. Thus, "beat quantity tables" published for the industry show several percent fewer beats.

Figure 10.10 is a plot of the number of beats for some common, fully loaded, system bandwidths.[9] As can be seen, channels slightly above the middle of the spectrum are affected by the greatest number of beats. It does not follow, however, that the channel with the greatest number of beats is the channel most affected by third-order interference. Amplifiers are not typically operated with equal power carriers, but rather with a uniform upward "slope" of levels with frequency (for reasons explained later in this chapter), causing some beats to be stronger than others. More important, since signal quality is affected by the *ratio* between visual carrier level and interference level, the lower carrier levels at the low end of the spectrum will sometimes cause the worst interference to occur on a lower-frequency channel than the one with the largest number of beats.

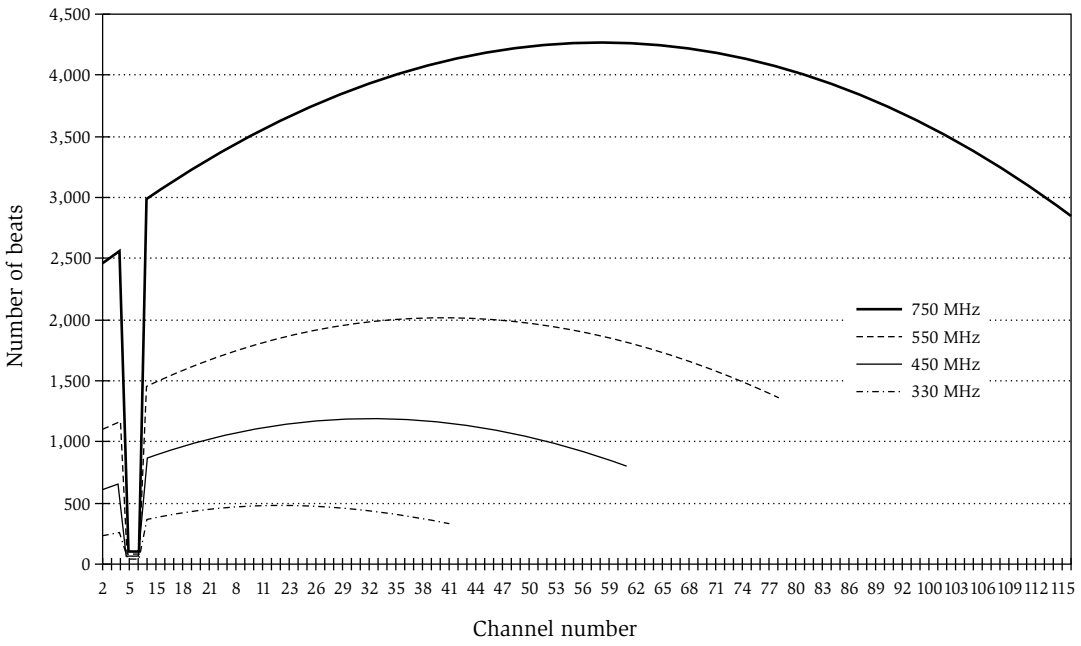


Figure 10.10 Number of triple beats per channel versus system bandwidth.

Since the third-order products are phase incoherent in most systems, their powers will add on a power, rather than a voltage, basis, and the measured CTB level would theoretically be proportional to $10 \log(n)$, where n is the number of beats in a channel if the levels of all carriers are equal. In actuality, owing to the unequal carrier levels and beat amplitudes, and possibly differences in distortion across the spectrum, CTB may increase as slowly as $5 \log(n)$ to $7 \log(n)$.[10]

The visual effect of CTB on an analog television picture depends on whether the signals are harmonically related or not. With unlocked carriers, the differences in frequency between the desired carrier and the various beats produces low-frequency interference that has been described as "streakiness" in the picture. With locked carriers, the various products will be frequency coherent with the carrier and, like the fundamental frequency products, will alter the amplitude and phase of the original signals. The visual effects depend on many factors, including the amount of phase noise on each of the signals.

Since analog video carrier modulation results in an average signal level that varies with program content, equipment is specified and evaluated when loaded with unmodulated carriers of a defined level on each of the nominal visual carrier frequencies. This results in more repeatable measurements. When loaded with video carriers whose peak level is the same as the unmodulated levels used for the test, the average carrier power will be about 6 dB lower, resulting in an improvement of about 12 dB in CTB (though the improvement will vary over time depending on video modulation levels and timing and carrier phase relationships). Typical amplifiers provide a CTB of -70 to -90 dBc when loaded with unmodulated carriers at recommended operating levels, depending on design.

Cross Modulation (XMOD)

Another form of distortion occurs when amplifiers exhibiting third-order distortion are loaded with multiple amplitude modulated signals.

Although second-order distortion has no products that affect the gain of the primary FDM carriers, as Equations (10.27), (10.29), and (10.30) demonstrate, two classes of third-order products affecting every channel cause effective gain variations of the fundamental signal. In particular, the first term of Equation (10.30) demonstrates that the effective amplifier gain affecting each FDM subcarrier is modified by a series of terms that include the amplitude of every other signal in the spectrum carried. Thus, when the average levels of any other signals are varied because of modulation, the output level of the desired channel will also vary slightly. This effect is known as cross modulation.

If the modulation of the other signals were truly independent of each other and random, the effects of cross modulation from various channels could be expected to be noiselike. Unfortunately, video modulation is not random. As covered in Chapter 2, the change in amplitude from synchronizing peak to black level is identical for every channel and, furthermore, occurs at nearly

identical and quite precise frequencies. Thus, if the signals modulating multiple channels happen to be synchronized (which happens when they are locked to a common timing reference in a multichannel generation facility), or if they drift into alignment occasionally, the level of cross modulation can become significant.

The standard test for cross modulation is to load the amplifier with a comb of simultaneously modulated carriers plus one unmodulated carrier on the frequency to be tested. The level of modulation is measured on the nominally unmodulated carrier. Cross modulation is defined as the difference between the cross modulation sideband level and the sideband level that would correspond to 100% modulation expressed in decibels.

Since it is caused by third-order distortion, the level of cross modulation varies in the same way as CTB; that is, the relative level varies 2 dB for every 1-dB change in operating levels. Because of the difficulty in making the test and demonstrated inconsistency in the results, cross modulation has become less important than CTB as a means of characterizing third-order distortion. The inconsistency may be due to the many terms, whose effects may reinforce or cancel, affecting the cross modulation on any channel.

When the desired signal is an analog television program and the major cross-modulating signals are also analog television programs, the first visible indication of cross modulation takes the form of a single, faint broad horizontal line and another broad vertical line that may slowly move through the desired picture. These are caused by the horizontal and vertical synchronizing pulses (the highest carrier levels) of another television signal. In an improperly operated system, cross modulation can also be caused by a single very high amplitude channel that causes cross modulation onto other channels. In extreme cases, not only synchronizing bars but video modulation may be visible. The visual appearance of cross modulation is similar to that which results from direct pickup interference where an undesired signal "leaks" into the system and interferes with a cable signal on some channel.

Composite Intermodulation Noise (CIN)

The preceding analysis was based on the carriage of a spectrum of signals, each of which has a carrier that represents most of the energy in the channel. Though that is a valid model for analog television signals and many conventional AM and narrow band FM modulated signals, digital signals differ significantly. A QAM signal, for instance, has a suppressed carrier and, when viewed on a spectrum analyzer, looks like a flat block of noise occupying the entire communications channel (see Chapters 4, 5, and 6 for detailed discussions of QAM modulation).

When such signals are subject to second- and third-order distortion, they do not produce single frequency products, but rather noiselike bands of

energy whose amplitude frequency signature depends on the distortion mechanism. The products of this type of distortion are alternatively known as composite intermodulation noise (CIN), composite intermodulation distortion (CID), or intermodulation noise (IMN). This is an exception to our use of the term *noise* to refer to only thermal noise since CIN is actually a distortion product.

On the assumption that third order is the dominate distortion-producing mechanism and that, further, three frequency products ($\omega_a \pm \omega_b \pm \omega_c$) dominate, IM products involving digital carriers will be of three types:

- Products formed from the mixture of two analog carriers and one digital signal will appear as blocks of flat noise with a bandwidth equal to that of the digital signal.
- Products formed from the mixture of one analog signal and two digital signals will appear to have a symmetrical triangular-shaped amplitude-versus-frequency spectrum, where the triangle will have a spectral width equal to the sum of the widths of the two individual digital signals.
- Products formed from the mixture of three digital signals will appear to have a band-limited Gaussian shape with a spectral width equal to the sum of the widths of the three mixing signals.

If digital signals are carried lower in level than analog signals, then the preceding types of products are listed in decreasing order of amplitude.

Although standard operating practices have not yet been set for the cable industry, it has been reported that a 750-MHz plant, operating with a signal load of 77 modulated analog television signals (54–550 MHz) and 33 64-QAM digital signals (550–750 MHz) and typical operating conditions and levels, will suffer less than 1 dB of effective C/N degradation due to CIM with digital signal levels suppressed as little as 5 dB relative to analog signals.[11]

10.3.4 Group Delay Variation and Amplifier Stability

Group delay is a measure of the rate of change of phase shift through a device as a function of frequency. In equation form

$$\text{group delay (sec)} = \frac{1}{2\pi}\frac{d\gamma}{df} \tag{10.35}$$

where

γ = the phase shift through the device in radians
f = the frequency in Hz

In an ideal RF system, the transmission delay at all frequencies is a constant equal to the transit time through the length of line involved, and the group delay is zero.

If the group delay is not zero, signal components at one frequency will arrive shifted in time relative to other components. The degree of tolerance to group delay variation will vary according to the signal format. A particular case of interest to the analog NTSC format is chroma delay, a measure of the arrival time of the chroma signal relative to the luminance signal. As discussed in earlier chapters, the difference should not exceed 170 ns. Though modulators and demodulators are generally the primary contributors to chroma delay, channels close to forward or reverse transmission system band edges are also affected by amplifier design.

High-speed digital signals are also affected by group delay. For example, fast rise time waveforms will broaden if the higher-frequency components are delayed relative to the lower-frequency components. Occasionally, digital signals are carried below channel 2 in the downstream band and, thus, very close to passband edge.

The principal contributors to group delay variation in amplifiers are the diplex filters used at each of the signal ports to separate upstream and downstream spectra. Figure 10.11 illustrates the key components of a two-way amplifier. The input and output diplex filters each consist of a high-pass section connecting the forward gain block to the external ports and a low-pass section feeding the reverse gain block. Each of the filter sections, as well as the amplifier modules, will have a gain (negative in the case of the filters) that is a function of frequency.

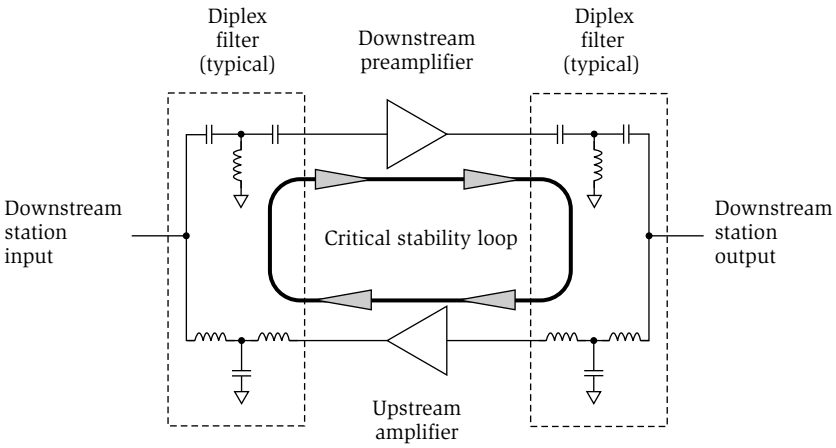


Figure 10.11 Bidirectional amplifier.

In order to assure absolute stability, the sum of the gains around the loop shown in the heavy arrow must be less than zero at all frequencies. This condition must be met not only for matched loads on the external ports but for any mismatch encountered in an actual field installation. With simple filter designs, this is most easily achieved by having the cutoff frequencies of the filters as far apart as possible and, thus, close to the edge of well-separated forward and reverse passband edges. Since group delay variation gets worse as the cutoff frequency is approached, there is an inherent conflict among (1) the separation between the top of the upstream spectrum and the bottom of the downstream spectrum, (2) the cost of the diplex filter, (3) the group delay variation at the extremes of the spectra, and (4) the stability margin of the amplifier.

Ensuring a nonoscillatory condition may not be adequate. Even if the loop gain is less than zero, frequency response variations may result unless there is adequate isolation. As a practical test, adequate performance may be deduced by observing the swept frequency response of an amplifier at all frequencies, including the diplexer crossover region. Peaking of the response indicates a possibility of a loop gain problem.

Figure 10.12 illustrates the downstream performance of a typical commercially available amplifier having a "42/52 MHz split" (upper edge of the upstream spectrum is at 42 MHz, whereas the lower edge of the downstream

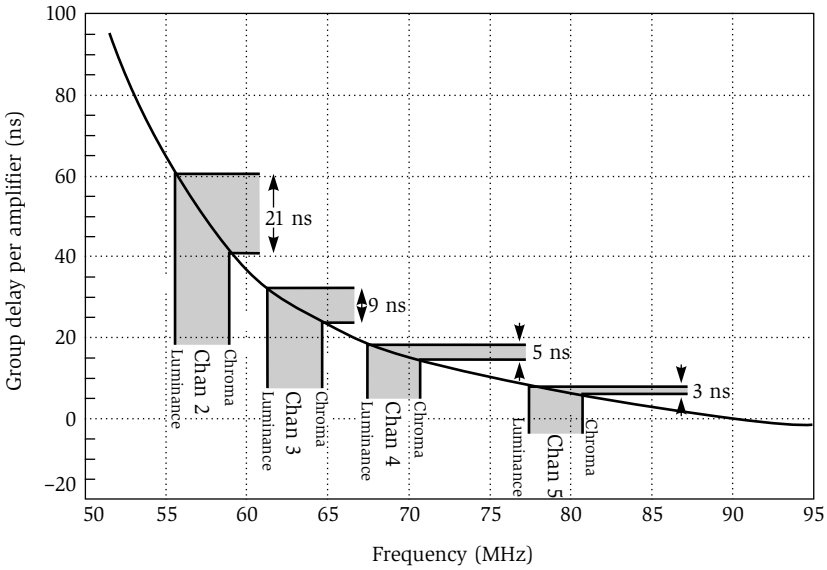


Figure 10.12 Downstream group delay due to diplex filters.

spectrum is at 52 MHz). The curve illustrates the group delay, with the shaded regions indicating the frequencies and chroma/luma delay difference for the four lowest NTSC television channels in the Standard ANSI/EIA-542 frequency plan. The direction of the chroma delay is to advance the chroma with respect to the luminance. Serendipitously, this is the opposite delay sense (luminance is delayed with respect to chrominance) to that which is commonly encountered in the modulation–demodulation process, but any delay is not good.

The required end-to-end group delay requirements and typical network performance are discussed in Chapter 15, and upstream group delay is discussed in Chapter 16.

10.3.5 Hum Modulation

Although amplifiers that plug into conventional wall outlets are available (and used in headend and apartment house applications where convenient), most distribution amplifiers are powered via ac voltage that is multiplexed with the signals on the common coaxial distribution cable. Along with their signal-processing circuits, amplifier stations include circuitry for separating the power and signal voltages and for connecting the power to local power circuits that convert it to direct current to power the various electronic modules used.

Amplitude modulation of the transmitted signals at the power line frequency (hum) can occur in two ways: excessive ac voltage at the output of an amplifier's local power pack and parametric modulation of magnetic component properties. Power pack "ripple" can modulate amplifier gain or be coupled into other circuits such as the automatic gain control (AGC). This ripple can result from faults in the power pack or loss of voltage regulation due to excessive voltage drop between the power supply and the amplifier's power pack.*

Parametric magnetic modulation is less obvious. Consider the circuit segment in Figure 10.13. The current is tapped off the common coaxial input and sent through L to the amplifier power pack (and possibly inserted into the output coaxial cable(s) as well). Series capacitor C serves to pass the signals into the amplifier while blocking the supply voltage. Finally, T is the input winding of a typical signal-processing element, such as a directional tap, splitter, or part of a band-splitting filter.

One possible source of hum modulation is L. This component is required to exhibit a very high impedance to frequencies covering the entire forward and return spectra. If the magnetic core becomes partially saturated at the peaks of the supply current, then the impedance match of the amplifier will vary

---

* In accordance with standard industry terminology, the term *power supply* is used to designate the device that generates the system power for multiplexing on the coaxial cable, whereas the term *power pack* is used to designate the device within each amplifier that converts the multiplexed power to the appropriate regulated dc levels for the circuitry.

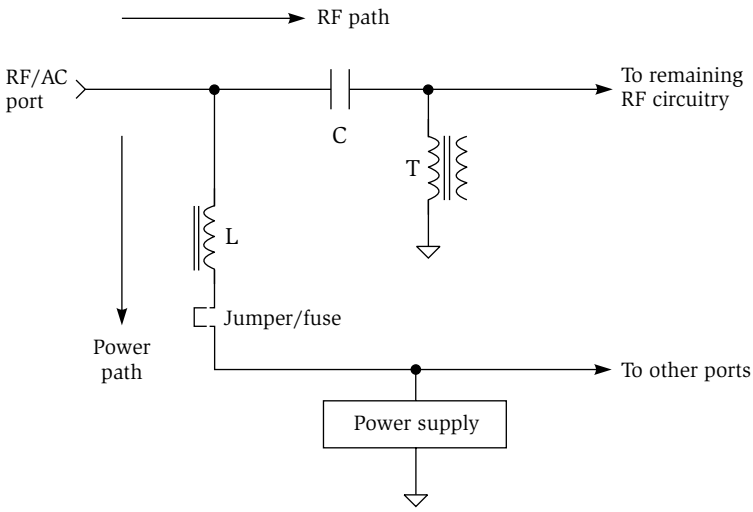


Figure 10.13 Power/RF separation in amplifiers.

accordingly, resulting in a level variation of the transmitted signal at the line frequency rate.

T is a much more sensitive component. Typically, these transformers are constructed of tiny cores that must have constant magnetic characteristics over an extremely wide frequency range. Though C nominally protects T from the supply current, there is still a displacement current to contend with. In particular, an ac current will flow through C and the primary winding of T whose magnitude is

$$i_C = \mathrm{C}\frac{dV}{dt} \tag{10.36}$$

When the power supply is a sine wave with a frequency of f_p and a peak voltage of E_p, the current will be

$$i_C = \mathrm{C}\frac{d}{dt}[E_p \sin(2\pi f_p t)] = \mathrm{C}E_p 2\pi f_p \cos(2\pi f_p t) \tag{10.37}$$

which has a peak value of $\mathrm{C}E_p 2\pi f_p$. If the voltage is a square wave with fast rise and fall times, then the displacement current can be much greater for the same peak supply voltage. To illustrate the trade-offs in design of these circuits, assume that it is desired that C have no more than 1 ohm of reactance at the low end of the upstream spectrum (generally 5 MHz) so as to affect the input match

as little as possible. This requires that C be at least 0.03 μF [$X_C = 1/(2\pi f C)$]. If the supply voltage is 90-volts rms (127 volts peak) at 60 Hz, then the displacement current will be 1.4 mA. If, under square wave powering conditions, the rise time is five times that under sine wave conditions, then the current could be as high as 7 mA. The designer's challenge is to ensure that this level of additional current in the winding of T does not change the transformer's characteristics.

In the North American NTSC analog television system, the frame repetition rate is 59.94 Hz. Since the commercial power frequency is 60 Hz, hum modulation generally appears as a horizontal bar or brightness variation whose pattern slowly moves upwards with time (about 16 seconds to move the entire height of the picture). Typical coaxial components (both amplifiers and passive elements) are specified to create hum modulation at no greater than −70 dB, relative to 100% modulation, at their full rated supply voltage and pass-through current.

10.3.6 Frequency Response Variation and Impedance Match

Practical amplifiers do not amplify signals at all frequencies equally. Part of this variation is intentional and is intended to compensate for variations in cable losses and to optimize distortion performance, as we discuss in the next section. Any deviation from the intentional variation, however, is known as frequency response variation (quantified as $\pm x$ dB) or, more commonly, peak-to-valley (quantified as y dB peak-to-peak). Response variation is defined by the SCTE as the maximum gain deviation in either direction from a line drawn from the gain at the low end of the defined bandwidth to the gain at the high end of the defined bandwidth.[12] Figure 10.14 illustrates this definition; the peak-to-valley (P/V) is the sum of the absolute values of $+\Delta$ max and $-\Delta$ max. This definition, however, is

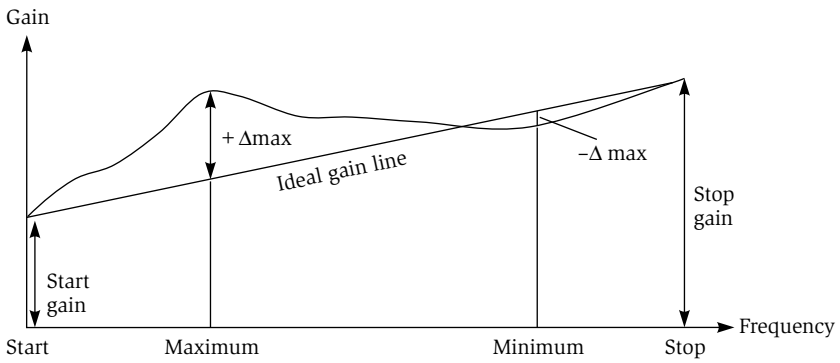


Figure 10.14 Definition of response variation.

somewhat controversial since it is highly affected by the endpoints. An alternative definition for frequency response is the variation above and below a line that is drawn so as to minimize P/V. For instance, if the lower end of the "ideal gain line" in the figure were raised slightly, the effect would be to lower $+\Delta$ max by more than the increase in $-\Delta$ max and thus reduce the total P/V.

The frequency response through a network may be affected by the imperfect impedance matching between the amplifier and attached transmission line. The effects of imperfect matches in coaxial components can best be understood in the context of their interaction. This is covered in detail in Chapter 15.

10.3.7 Practical Amplifier Design Choices

Complete amplifier stations may be simple or quite elaborate, depending on system needs. All share a common basic downstream structure consisting of several (generally two or three) cascaded gain stages, plus circuits that compensate for the loss variation of the cable before the amplifier (equalizers). Provisions are also made for adjusting the gain of the amplifier as needed to meet system needs. Amplifiers differ in how many stages are employed, internal splitting, redundancy provisions, where operating adjustments are in the chain, automatic gain control, and the performance of each gain block. Many have provisions for automatically adjusting amplifier gain to compensate for changing input signal levels (automatic gain control).

As will be seen, optimum operation of repeatered coaxial distribution systems results when the levels of the FDM subcarriers at the output of amplifiers are not equal but increase with frequency. The variation in level across the operating range is known as *tilt*. The relationship between operating levels, amplifier response, and interconnecting transmission losses is explored more fully in Section 10.3.8.

In the upstream direction, all the amplifier designs are similar. Since cable losses are much lower in that portion of the spectrum, a single amplifier module is sufficient. Diplexers are used at each port to separate the upstream and downstream signals. All the upstream signals are passively combined and fed to the amplifier module. Unlike downstream, where all level setting takes place at the input or midstage with the aim of setting a consistent *output* level, the common upstream practice is to align the system so that the *input* level to the module is constant and to set the output gain and slope after the gain stage so that the input to the next amplifier station upstream is correct. Chapter 16 covers the alignment and operation of networks in the upstream direction.

To allow the most flexibility in application and system design, jumpers or switches allow power to enter the amplifier via any desired port and to be routed to any combination of output ports.

Early power packs used conventional transformer/linear regulator designs. Transformer taps were available to optimize the winding ratio of the transformer to limit required regulator dissipation. Modern designs almost universally use switching regulators (whose design is beyond the scope of this book) that have the property of operating at a constant high efficiency regardless of input voltage. To do so, they draw a constant *wattage* from the coaxial cable, and thus the current is inversely proportional to the supply voltage. Though this results in the greatest overall powering efficiency, it has some unexpected consequences when a number of amplifiers are powered from a common source, as will be seen in Chapter 11.

Beyond their basic functional elements, amplifier stations may be equipped with a variety of enhancements, including redundant gain modules and power supplies, status monitors and transducers, and remotely controllable upstream path disconnect. Often fiber-optic receivers and transmitters are combined with amplifier station components, creating a fiber-optic node.

Although the labels attached to different configurations are somewhat arbitrary and subject to change, the following classifications are typical:

- Trunk amplifiers are the workhorses of traditional all-coaxial cable television systems. They are optimized for long-distance, repeatered signal distribution. As such, they are operated at output levels that cause only moderate distortion to allow long cascades. The gain is set as high as possible, consistent with low per-stage noise addition. Typically, the output levels of the highest channels are +35 to +40 dBmV, and the gain is about 22 dB. Trunk amplifiers are not used to supply signals to taps.
- Where a local distribution circuit is also desired, a portion of the trunk signal is fed to a separate output stage known as a bridger. This amplifier generally runs at a higher output level (and distortion) to allow feeding as many customer taps as possible. Though generally configured as an optional module within the trunk station, bridgers may be contained in separate housings along with directional couplers (midspan bridgers). Figure 10.15 is a block diagram of the RF portion of a typical trunk station, including bridger.
- Distribution amplifiers are basically stripped-down trunk/bridger stations that are available in a wide variety of configurations. Typically, they are available with multiple output amplifier modules whose level can be independently adjusted. Figure 10.16 shows a typical configuration with three output gain blocks.
- Line extenders are used to allow feeding more taps than possible with just a bridger or high-level distribution amplifier port. Since they are generally operated at similar output levels, the cascade of such devices must be limited to keep distortions at an acceptable level. Line extenders are the simplest of the classifications, as shown in Figure 10.17.

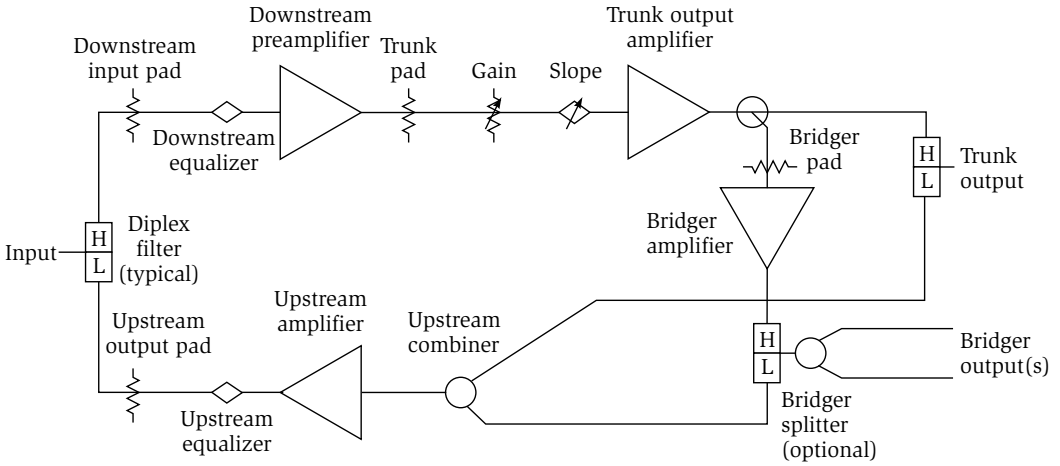


Figure 10.15 Typical trunk amplifier RF configuration.

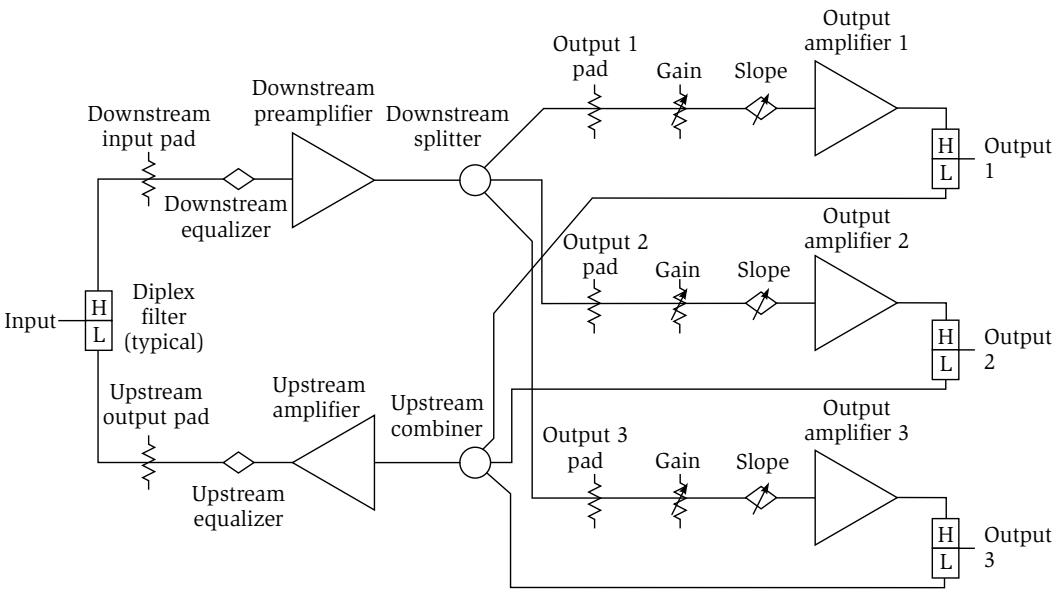


Figure 10.16 Typical distribution amplifier RF configuration.

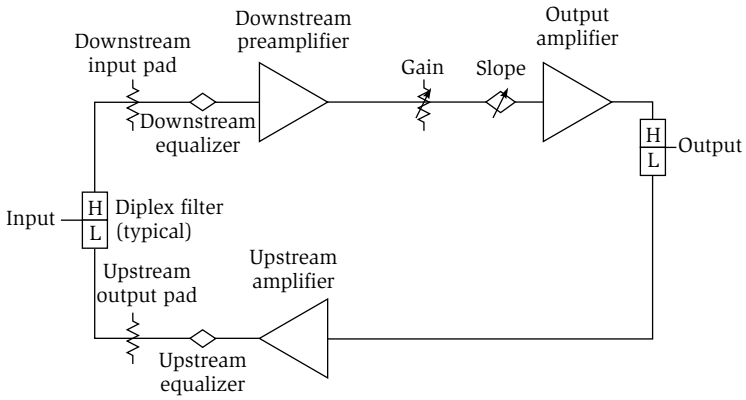


Figure 10.17 Typical line extender RF configuration.

10.3.8 Amplifier Operating Dynamics

The C/N degradation as a signal passes through an amplifier is related to the input signal level and the noise figure at any frequency (see Equation (10.18)). On the other hand, the loss of coaxial cable increases with frequency. Thus, the amplifier need not supply the same level for channels at the bottom of the spectrum as at the top to overcome cable loss. Reducing the levels of the lower-frequency channels, in turn, reduces the amplitudes of the intermodulation products to which those channels contribute, and thus CSO, CTB, CIN, and XMOD.

The relation between cable and amplifier characteristics and operating levels can best be illustrated by considering an amplifier and its driven cable as a system (see Figure 10.18). Only the downstream path will be discussed here; Chapter 16 will deal with the equivalent upstream issues.

The preamplifier will typically have a relatively uniform noise figure and gain across the spectrum. Between the preamplifier and output stage, a slope circuit is used to adjust the frequency response so that it uniformly increases across the downstream frequency range. Typical gain slopes are 6–10 dB for a 750-MHz amplifier. Additionally, pads are used to adjust the overall amplifier gain. The station noise figure and slope are illustrated in Figure 10.19(a).

Some combination of coaxial cable and passive RF devices is connected between this amplifier station and the input of the next station. Since both the cable and passive devices have a loss that varies across the spectrum (and, in the case of cable, increases as the square root of frequency), the variation in level reaching the next station will be reduced. Amplifier input circuits are designed to accommodate a variety of equalizers that are optimized to

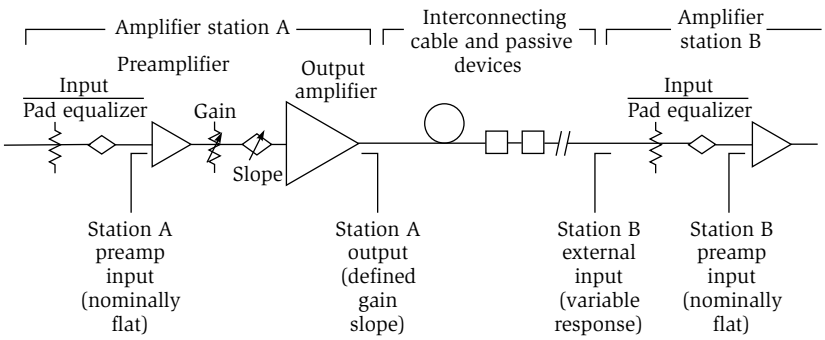


Figure 10.18 Downstream response and gain relationships and adjustments.

compensate for the residual response variation and pads that are used to set the correct drive level to the preamp input. Occasionally, external in-line equalizers may be used also, for example, as a tool to reduce the total variation of levels fed to customers.

In station alignment, the equalizers and pads are used to align the *output* of each amplifier station as closely as possible to the optimum levels and "gain slope." Thus, the system gain and frequency response, as measured from the output of one amplifier to the output of any other similarly aligned amplifier in a series-connected string, are ideally unity. Because of the variability in interstation losses, the levels and response at station inputs will vary; however, the response at the internal preamp input will be approximately flat. Amplifiers that do not feed customer taps (for example, trunk amplifiers) are typically operated at levels several decibels lower than those that do.

By contrast to the frequency response of the system (slope), the variation in levels of the signals carried is known as tilt. The terms are often confused, particularly since modern amplifier performance is often specified with a tilt at amplifier outputs that matches the gain slope. Historically (with narrower bandwidth amplifiers), it was more common to use "block tilt" with groups of channels operated at different levels. This was easier for technicians to understand since only a few levels needed to be memorized.

Modern 750-MHz amplifiers are generally specified under two signal-loading conditions. Figure 10.19(b) shows the nominal preamplifier input (after equalization) and output levels of an amplifier station loaded with a full spectrum of 110 video signals that are tilted to match the gain slope of the amplifier. This is the optimum condition for carrying signals of similar format and results in approximately equal C/N degradation for each channel (because of the uniform preamp input levels) and the lowest distortion (because the outputs at the lower channels are reduced in level to approximately match the interconnecting cable

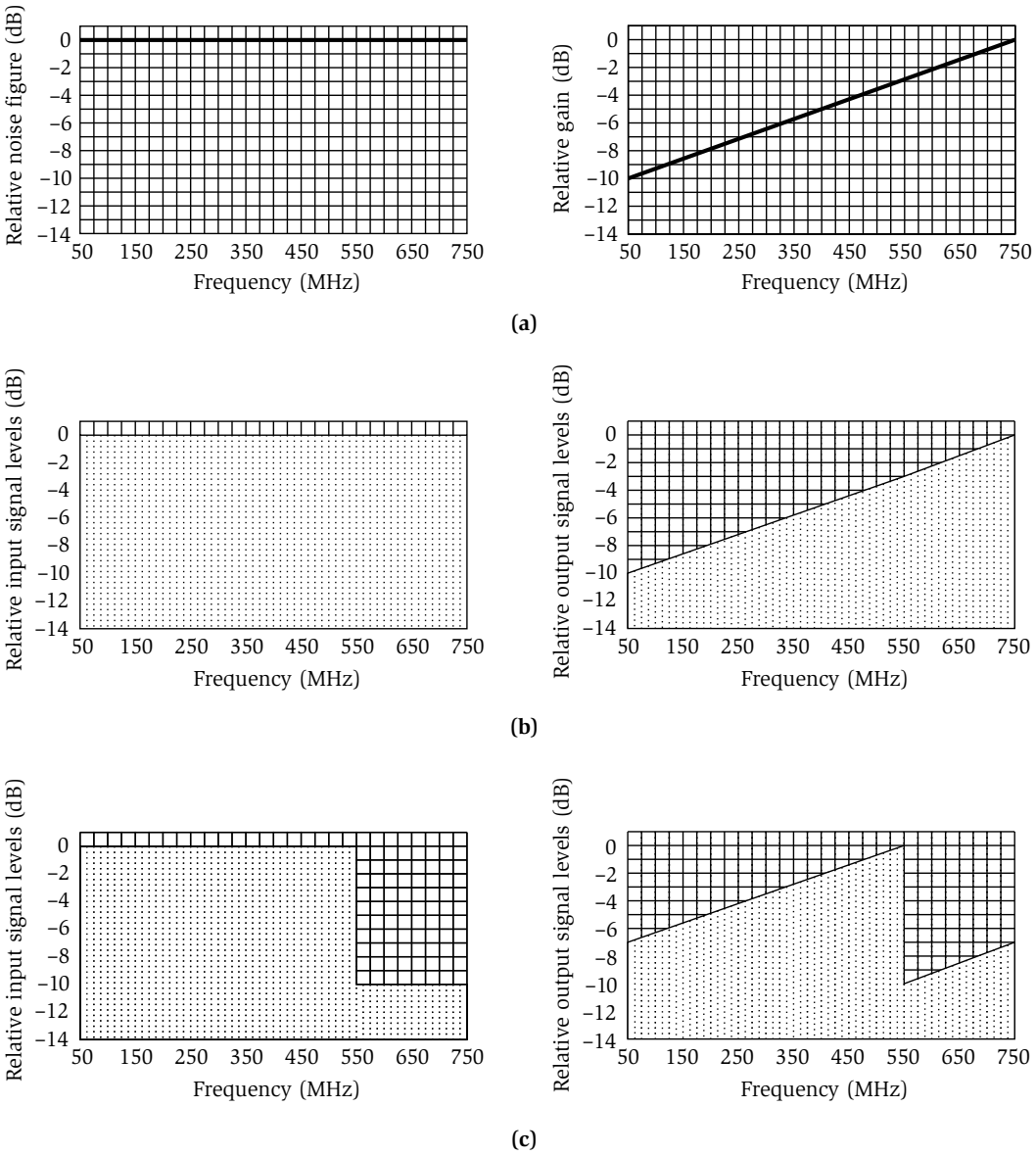


Figure 10.19 Amplifier gain and signal loading. (a) Amplifier characteristics. (b) Amplifier loaded with analog video signals. (c) Amplifier with hybrid analog–digital signal loading.

and component losses). Figure 10.19(c) shows an alternative loading condition with a superimposed −10-dB block tilt for signals above 550 MHz. The assumption is made that these signals are digitally modulated and do not require the same C/N and C/distortion as analog video channels and so are reduced in level to reduce overall amplifier loading.

Optimum amplifier gain is a function of noise figure and output capability (for a given distortion level). By way of illustration, assume that a given amplifier has an effective noise figure of 9 dB and a CTB of −78 dB at an output level of +45/35 dBmV (the dual numbers are the individual channel levels at the top and bottom of the downstream spectrum, respectively). If the gain of the unit is set at 35 dB (at the top channel frequency) and the input level is +10 dBmV on all channels, then, from Equation (10.19), the C/N of this stage alone will be 60 dB (10 + 59 − 9). The output level will be the input level plus the gain or +10 dBmV +35 dB = +45 dBmV, so the CTB will be the specified −78 dBmV.

If the input level is varied, the C/N will improve by 1 dB for every decibel increase in input level, whereas the CTB will degrade by 2 dB. In setting up a coaxial system, input padding is used to set the optimum operating point.

Now consider the situation if the amplifier had been designed with a gain of 40 dB. At first glance, this would seem like an improvement since such amplifiers could be spaced farther apart along the transmission line. When driven by a +10 dBmV input signal, however, the output level will be +10 dBmV+40 dB = +50 dBmV. At that level, the CTB will be only −68 dB, having degraded by 10 dB owing to the 5-dB increase in output level. If the input level is decreased by 5 dB to restore the distortions to the specified level, the station's C/N will decrease from 60 to 55 dB. In short, the excess gain forces the user to accept either higher distortion or higher noise.

Though line extender amplifiers often offer only an input padding option, some distribution and trunk amplifier designs also allow between-stage padding. This partially offsets potential excess gain problems by allowing the user to independently choose the operating level of each gain block though each gain block is still limited in performance by its individual noise figure, gain, and output capability.

The selection of optimum amplifier internal gains is very complex. It would seem from the preceding example that reducing gain would improve performance. Although that is true, more lower-gain amplifiers would have to be cascaded to reach the same distance, and that would lower reliability and increase power consumption. The industry's choice of operational gains ranging from 20 to 38 dB (depending on technology and application) is based on the best compromise among cost, reliability, and performance.

10.3.9 Amplifier Technology Choices

As previously discussed, modern broadband solid-state cable television amplifiers are all push–pull in nature, meaning that their output stages contain two

matched transistors connected in a symmetrical configuration. To the extent that any nonlinearities are also symmetrical, they will produce only odd-order distortion products. Amplifiers whose output stages each contain only a single push–pull circuit are known generally as *push–pull amplifiers*.

An improvement in distortion performance can be realized by splitting the input signal, feeding two push–pull output gain stages in parallel, and then combining the outputs into a single port. This type of amplifier, generically known as *parallel hybrid*, offers about twice the output level at the same distortion levels compared with push–pull units. Since two output stages are involved, the power consumption and dissipation are also higher than with push–pull units. Parallel hybrid amplifiers are generally used where high output powers are beneficial, such as bridgers and line extenders.

The third dominant amplifier technology is known as *feed forward*. In a feed-forward amplifier, the input signals are amplified via a conventional inverting amplifier. Samples of the input and output signals, adjusted for gain, are combined and amplified in a separate inverting *error amplifier*. The output of the error amplifier is summed with the output of the original amplifier so that the errors approximately cancel.[13] Figure 10.20 shows this configuration. The delay lines are required to match the phases of the signals. Modern feed-forward amplifiers offer about 12 dB of CTB improvement and 6 dB of CSO improvement over power doubling units at comparable operating levels. Furthermore, as discussed later, the remaining distortions are not necessarily phase coherent among cascaded amplifiers, which improves the performance of a repeatered system. Feed-forward amplifiers are generally used in long trunk lines where the lower cascaded distortion is beneficial.

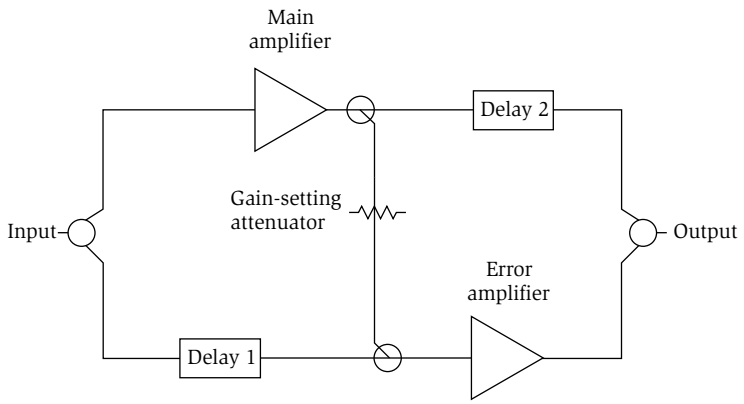


Figure 10.20 Feed-forward gain block.

10.4 Passive Coaxial Components

The nature of typical coaxial distribution systems is that the downstream path is split and resplit to create many endpoints from a common signal insertion point. In the upstream direction, the opposite is true, with signals from many potential insertion points combined at a common node. In the headend, signals from individual modulators and signal processors are combined to create the full downstream spectrum, whereas upstream signals are split to feed receivers for each service.

The signal splitting can occur either within amplifier housings or by use of stand-alone components. Those that split the signal equally are known as *splitters*, whereas those that divert a defined portion of the input signal to a side port are known as *directional couplers*. Finally, those that divert a portion of the input signal, then resplit the diverted signal to create individual customer drops are known as *taps*. All share certain characteristics and so will be treated in this section.

In general, all these are passive devices constructed using ferrite-loaded transmission lines and transmission-line transformers. They are bidirectional; that is, the same device can be used to split signals to feed multiple paths or to combine signals from multiple input ports. Important characteristics are signal loss, impedance match, and isolation between nominally isolated ports.

10.4.1 Directional Couplers

Figure 10.21 shows schematics of two directional coupler configurations. In both examples shown, the voltage transformer and current transformers have the same ratio so that, after combining, the ratio of voltage to current stays the same, meaning that the impedance of the side port is the same as the main line. In the main line, the relative polarity of voltage and current is reversed for signals traveling in one direction versus the other direction. The transformers sample voltage and current regardless of signal flow in the main line. The difference is that the sampled signal, after combining, will propagate either toward the termination resistor or toward the side port, depending on the relative polarity of voltage and current. This is what gives the coupler its directivity.

By varying the turns ratio of both transformers, various values of coupling are possible. Ellis[14] has published a tabulation of coupling values as a function of turns ratios, along with a good description of coupler operation. Table 10.2 lists some commonly used values and their theoretical signal loss values (practical units will exhibit 1–2 dB of excess loss due to imperfect components). The coupled loss is simply due to the voltage (or current) ratio defined by the transformers, whereas the input/output loss just reflects the loss diverted to the coupled port (or isolation resistor, depending on signal flow).

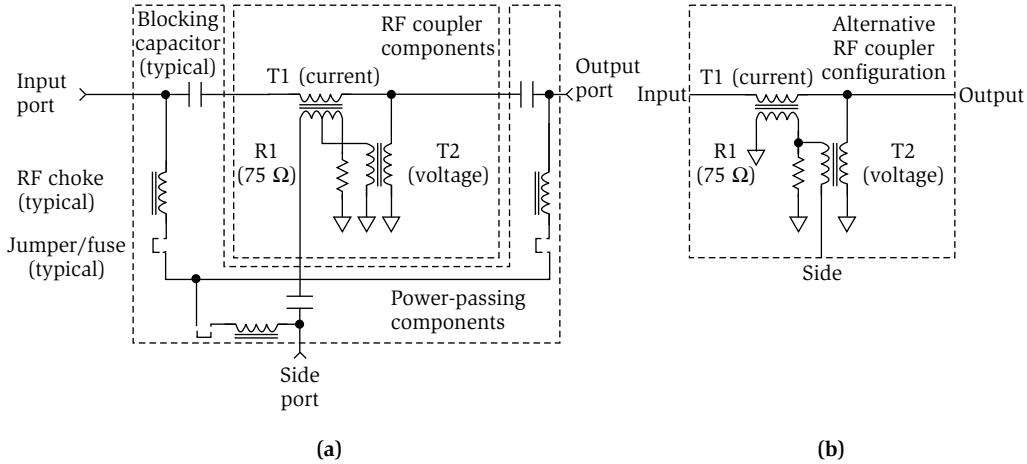


Figure 10.21 Directional coupler schematic. (a) Coupler plus power components. (b) Alternate RF configuration.

Table 10.2 Theoretical directional coupler losses

Turns Ratio (T1 and T2)	*Input to Output Insertion Loss (dB)*	*Input to Coupled Port Loss (dB)*
2.5:1	0.78	7.96
3:1	0.51	9.54
4:1	0.28	12.04
5:1	0.18	13.97

As is the case with amplifiers, added components in units intended for installation in trunk and distribution cables separate power from signal voltages and send them through different paths, as shown in Figure 10.21(a). Generally, means are provided to control transmission of power independently from RF signals so as to allow greatest flexibility in system design. Note that, if ac power is transmitted through all RF ports, the displacement current through the three blocking capacitors will add in phase through the primary of the voltage transformer. Thus, couplers generate parametric type hum modulation at levels as high as or higher than amplifiers.

The difference in sensitivity to downstream and upstream signals, as measured at the side port, is the *directivity* of the coupler. Sometimes, instead of directivity, the absolute upstream coupling factor, known as *isolation*, is specified. Numerically, directivity is simply the difference between coupling

factor and isolation. As we will discuss in Chapter 15, directivity and return loss interact to cause additional system group delay and response variation.

The coupling value may not be uniform across the entire frequency range of the device. Even though the overall variation in coupling factor may not be desirable or intentional, the slope and response variation are measured separately and are defined for passive devices the same as for amplifiers.

10.4.2 Splitters

Two-way splitters differ functionally from directional couplers only in that the signal splitting is equal. Other common configurations include four-way and eight-way equal splits. Three-way unequal splits are internally constructed by splitting two ways, then resplitting one of the legs from the first splitter to create a 50%:25%:25% output ratio. Equal ratio three-way splitters are less common.

Figure 10.22 illustrates a two-way symmetrical splitter, with expansion to four way. Figure 10.22(a) is the schematic diagram of a two-way unit, and Figure 10.22(b) is the standard symbol, with terminology. Transformer T1 is a 2:1 impedance step-down transformer that creates a 37.5-ohm source to feed the center tap of transformer T2. T2 is a transmission-line transformer derived from a Wilkinson coupler (which is used for symmetrical splitting of high-power, narrow band signals for directional antenna arrays, among other applications).[15] The two outputs at ports 2 and 3 are nominally in phase and at the same level. Resistor R1 is required to provide isolation between them.

Because power is equally divided between ports 2 and 3, the theoretical loss from port 1 to either output is $10\log(0.5) = -3.01$ dB. As with directional couplers, the practical loss will be higher, due to core losses in the ferrite core material in the transformers and resistive losses in the transformer wire.

The dimensions of transformers used in high-frequency applications are constrained by the necessity for the winding lengths to be a small portion of a wavelength at the highest frequency of operation. This, in turn, mandates very small cores and fine-wire gauges whose resistance is a factor in excess signal loss. The excess loss over and above that due to coupling is typically 0.5 to 1 dB to each of the output ports, making the total loss typically 3.5 to 4 dB.

Loss in a splitter is called *flat loss* because it tends not to be a strong function of frequency, as is the loss in coaxial cable. However, there is some increased loss as the frequency increases, partially due to skin effects, which increase the effective resistance of the wire as frequency increases. At the higher frequencies, the ferrite core on which the transformer is wound ceases to have significant magnetic properties, but rather acts as a coil form. Some loss may be associated with it, but the ferrite material is primarily included to improve low-frequency performance. At the lowest frequencies, loss also increases because the open-circuit reactance inductance of the transformer windings is no longer large relative to the circuit impedance.

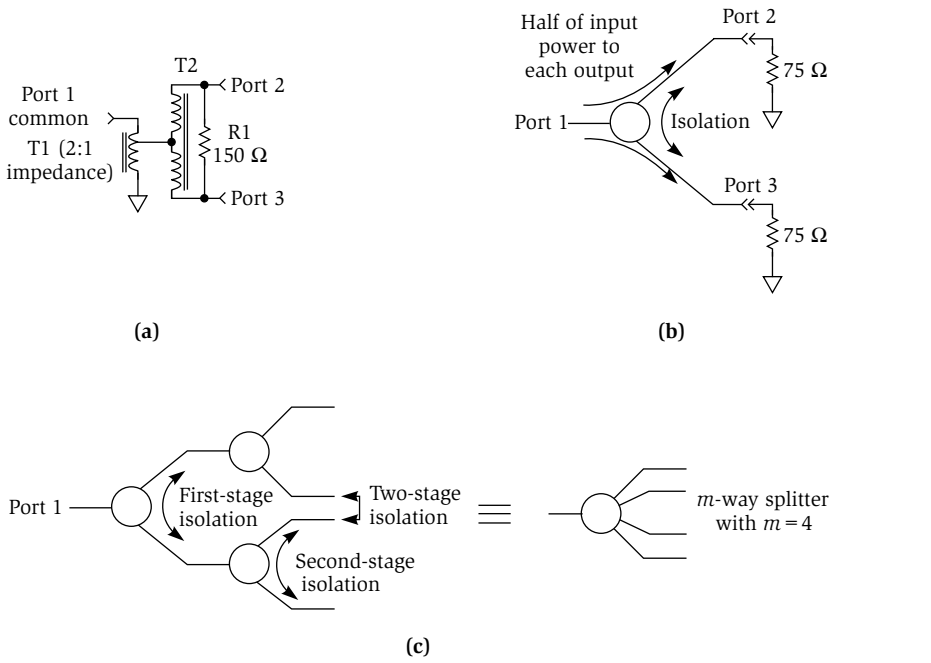


Figure 10.22 Splitters. (a) Schematic. (b) External power flow. (c) Multiple stages of splitting to create *m*-way splitter with enhanced isolation among some ports.

When the splitter is used as a combiner, incoming signals on ports 2 and 3 are combined and appear at port 1. The theoretical loss from port 2 (or port 3) to port 1 is also 3 dB, the other half of the power being dissipated in resistor R1. Theoretically, no power from port 2 appears at port 3 (if all ports are properly terminated) and vice versa. The signal loss between the output ports, with all ports terminated in 75 ohms, is defined as the isolation. Typical values of isolation for commercially available splitters range from 20 to 30 dB.

Other important specifications of splitters include insertion loss, normally specified as the total loss from port 1 to port 2, with port 3 terminated (or vice versa), and the return loss at each port, with the other ports terminated.

Lack of termination drastically reduces the effective isolation. For instance, power injected into port 3 is attenuated −3 dB at port 1. Should that port be unterminated, the power would be reflected into the splitter dividing equally between ports 2 and 3. Thus, the isolation from port 3 to port 2 would be only 6 dB. Similarly, the effective return loss at port 3 would be 6 dB as a result of power being reflected to that port.

Splitters having more output ports can be fabricated by combining two-way splitters in tree fashion, as shown in Figure 10.22(c). Each leg of the leftmost

splitter is connected to the input of another two-way splitter. At that level, four outputs exist. Each of those four can be connected to another two-way splitter, resulting in eight outputs, and so on. The practical loss at each stage will be 3.5–4 dB so that the four-way split loss is 7–8 dB. In an eight-way splitter, it is 10.5–12 dB and so on. In general, the loss in a multilevel splitter is given by

$$L_m = \left[\frac{L_1}{0.301}\right] \log(m) \tag{10.38}$$

where

L_m = the loss of the m-output splitter chain in decibels
L_1 = the loss in one two-way splitter (3.5−4 dB usually)

It is also possible to construct a splitter chain with a number of outputs not equal to 2^n by simply terminating the splits early on some of the chains. Equation (10.38) must be modified in this case.

Where output port isolation is important in multiport applications (such as headend combining) one alternative is to use ports in multilevel splitters that are not derived from common stages for critical applications. As shown in Figure 10.22(c), if two splitter outputs are connected to the same second-stage splitter, the isolation between them is merely the isolation of that stage. On the other hand, if two ports are not connected to the same second-stage splitter, then the isolation realized will include the effects of the first-stage splitter, as shown in the figure. Unfortunately, manufacturers may not indicate internal splitter configurations or specify nonadjacent port isolation.

Another higher-isolation alternative is to construct multiport splitters from cascaded directional couplers. Generally, side ports of adjacent, series-connected couplers will be better isolated than outputs from single-stage splitters.

10.4.3 Taps

The most common tap configuration consists of a directional coupler with the side arm signal split two, four, or eight ways to create customer drops. Although the input and through ports are typically provided with connectors appropriate for distribution cable, the side arms are equipped with connectors appropriate for drop cable. In North American systems, this means "5/8-24" ports for input and through ports and type F connectors for the drops.

A tap need not have a through port. At the end of the line, it is common to use a *terminating tap*, which is just a splitter whose output legs are equipped with drop connectors.

Until about 1995, taps supplied RF signals, but not power, to customer drop lines and were limited in main-line current-handling capacity to 7 amperes or

less since this was adequate for entertainment-based cable television systems. With the active consideration of HFC architectures for telephony service came increased demands on tap capabilities:

- Bandwidth expansion to 5–1,000 MHz
- The ability to supply power to side-of-house active *network interface devices* (NIDs) that could provide an interface between baseband twisted pair in-home telephone circuits and the RF carriers in the distribution system
- Provisions for changing tap values and configurations without interrupting service to customers farther along the distribution cable
- Increased voltage and current capability, typically 90 Vrms and 10 amperes, in order to allow transmission of sufficient power for both network amplifiers and NIDs over the same coaxial cables formerly used only for video service

The new generation of taps that have evolved are known as *power-passing taps* or, sometimes, *telephony taps*. In addition to normal tap capabilities, they tap into the distribution cable power and selectively provide current-limited power along with RF signals to a NID on each home. Depending on the design, this power may be multiplexed with the signals in the coaxial drop cable or carried along a separate twisted pair of copper wires in a hybrid cable configuration. In general, these taps also provide a make-before-break connection between the input and through ports when the tap "plate" (which contains all the circuitry) is removed from the housing so that both RF and power continuity are maintained to customers downstream.

A major concern for those using the drop center conductor to carry power to NIDs is that, if arcing occurs because of a faulty center conductor contact, the arc will transfer a very significant amount of RF power to the upstream plant, likely causing interference with all users of the reverse spectrum. On the other hand, an arc often causes healing of a bad contact and, thus, can be self-healing.

Figure 10.23 is a circuit diagram for a typical power-passing tap that uses the drop center conductor to pass power to the NID. The RF portion of the tap consists of a directional coupler and two-way splitter (the diagram is for a two-port tap). The directional coupler diverts a controlled portion of the downstream signal and passes it to the splitter, which divides the power to feed the subscriber ports.

Multiplexed ac power consideration in taps is similar to that on amplifiers and other network passives. Capacitors C1 and C2 isolate the multiplexed power supply voltage and current from the directional coupler RF components. Transformers T1 and T2 are built with very fine wire to permit operation at frequencies up to 1 GHz and would be unable to withstand the currents that would result if the primary of T1 were required to handle several amperes of ac current or if the primary of T2 had tens of volts of 60 Hz across it. The multiplexed ac power is shunted across the tap through the large RF choke L1.

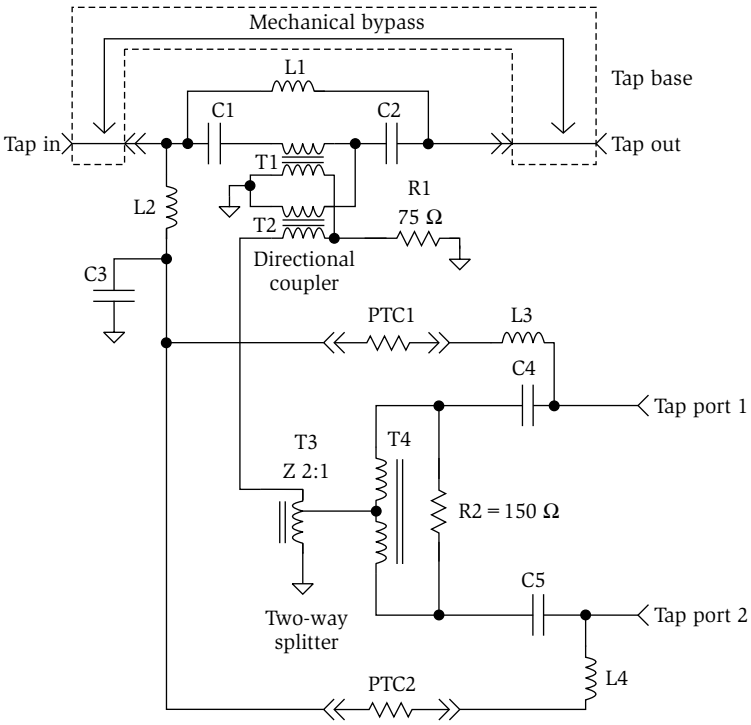


Figure 10.23 Power-passing tap using drop center conductor for power delivery to home.

A choke with somewhat less current capacity, L2, provides a path for the multiplexed power that will feed the customer ports. Since this choke is shunt-connected across the main RF path, it must have a higher RF impedance than L1. The unavoidable RF loss due to this choke is one reason power-passing taps have a slightly higher RF loss than normal taps.

In order to permit the tap plate (which contains most, if not all, the circuitry) to be removed without interrupting either power or RF continuity in the hard cable distribution plant, many modern taps incorporate a mechanical bypass switch that engages before the tap-to-baseplate connections are broken to provide continuity from input to through ports. During the removal or insertion of a tap plate, therefore, there will be a brief interval when both the bypass and main connections are made. This will cause a transient condition during which the primary of T1 will be shorted, but the primary of T2 will function normally. This momentary mismatch will cause some attenuation of the transmitted signals. Whether or not it causes a functional problem to the services offered will depend on the characteristics of the signals used.

Continuing with the description of the circuit, capacitor C3 is used to ensure a good RF ground on the "cold" end of L2. Positive temperature coefficient (PTC) resistors PTC1 and PTC2 limit the ac current that can be drawn through the customer ports. At normal currents, the PTC resistor (commonly called just a PTC) exhibits low resistance. As more current is drawn, its temperature increases, causing a resistance increase and effectively limiting the total current. Electronic current limiters have been proposed as an alternative to PTCs, but to date have not been widely deployed because of cost. New National Electric Code (NEC) regulations, effective in 1999, require a more sophisticated current-limiting system if the drop cable does not meet certain construction and installation criteria.

Finally, inductors L3 and L4 couple power to the two tap ports shown, again isolating the RF signals from the ac supply. As with L2, the reactance of these components must be sufficient to minimize RF losses, and their presence leads to slightly higher overall RF losses in commercially available power-passing taps.

The relationship between passive specifications and system performance will be explored in Chapter 15.

10.5 Power Supplies

Power supplies convert commercial power to lower voltages that are multiplexed with RF signals in the coaxial cables. Although not part of the RF circuit, certain characteristics of power supplies have an effect on the transmission of signals.

Direct current, very low frequency (as low as 1 Hz) ac, and 60-Hz ac have all been considered for system power, as have voltages between 30 and 90 volts. Direct current (dc) has been rejected primarily because of possible galvanic corrosion. Early cable television systems used 30 volts, but the increased power demands of wider-bandwidth, higher-output amplifiers forced the industry to convert to 60 volts for systems constructed after about 1970. Two factors are forcing consideration of even higher voltages: the desire to power larger plant sections from a single power supply to improve reliability (see Chapter 20 on HFC reliability for a detailed explanation) and the possibility of powering some terminal equipment through the distribution plant.

Along with the increase in voltage has come a requirement to increase currents as far as practical, with requirements to pass 10–15 amperes being common. Since current from supplies can be transmitted both ways along the cable and/or inserted into more than one cable, power supplies with 40 or more ampere capacity are available.

Field power supplies are roughly divided among nonstandby, standby, and UPS/standby types. The basic component of a nonstandby, line frequency power supply is the ferroresonant transformer. Though its magnetic design is beyond the scope of this book, ferroresonant transformers have a number of properties of interest to operators: (1) output voltage regulation (a typical specification is

±5% output voltage change when the input voltage changes ±15%), (2) inherent current-limiting and short-circuit protection, (3) transient protection, and (4) 90% or better efficiency at rated output current.

Standby power supplies add a battery charger, a set of rechargeable batteries, an inverter, and a changeover switch. The charger assures that the batteries are always held at full charge, whereas the inverter starts and supplies square wave power to replace the sine wave commercial power when it is interrupted (see Figure 10.24). Depending on the application, sufficient battery capacity may be provided for 2–8 hours or more of standby operation.

An important parameter for operation in the standby mode is the rise and fall time of the square wave voltage. As Equation (10.35) shows, the amount of ac power supply current that "leaks" into the low-level signal path in amplifiers and passives is directly proportional to the rise time of this waveform. The trade-off, for power supply designers, is that the efficiency of an inverter is an inverse function of the time the transistors spend in their active region (that is, not in a saturated or cutoff condition). Also, the power lost due to slow switching is dissipated in the transistors and must be conducted away by heat sinks. Unfortunately, the industry does not have a consensus on the best compromise on rise time.

When commercial power fails, there is an unavoidable but brief interruption while the relay switches between the ferroresonant transformer output and the inverter transformer output. Typically, the time to switch is 8–16 ms. Most supplies synchronize their inverters with commercial power before switching so that phase shift transients in the power supplies of the amplifiers can be avoided.

As will be discussed in Chapter 11, even with inverter phase synchronization, the short switching transient can be a problem when analyzed with respect to its effect on cascaded amplifiers. To avoid that, a third alternative is the uninterruptible power supply (UPS). Some UPS versions of field supplies avoid switching transients by utilizing the same ferroresonant transformer for both inverter and commercial power. During ac power outages, commercial power is disconnected, and the inverter continues to supply power via an additional primary winding.

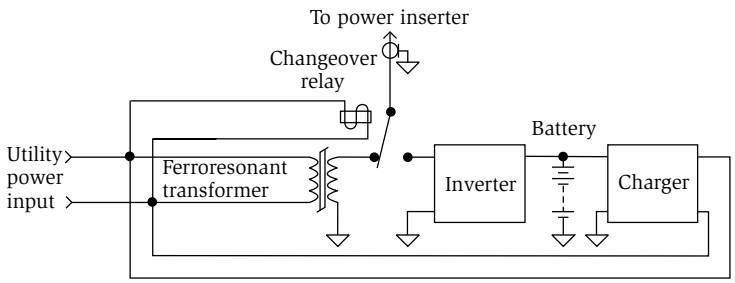


Figure 10.24 Basic standby power supply functional diagram.

10.6 Summary

This chapter has covered the technology behind coaxial distribution systems. The characteristics of the coaxial cable, amplifiers, passive components, and powering systems have been covered in detail, including both the theory behind operation and the performance of commercially available products.

Although other constructions are sometimes used, solid aluminum sheathed cable of 75-ohm impedance has been seen to offer a good combination of loss, cost-effectiveness, and handling characteristics and are in common use worldwide. Amplifier stations designed for cable systems are designed with unique characteristics that complement the characteristics of the interconnecting cable and passive devices. Integrated powering of active devices through the cable avoids the necessity of separately connecting each distribution system amplifier to the local utility, but requires special circuitry in every device to separate the power and signal paths.

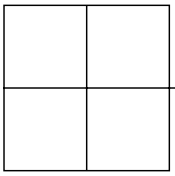
Before the mid-1980s, cable television systems used primarily coaxial technology (sometimes supplemented by broadband microwave links) to serve their customers. Chapter 11 will cover the design of coaxial distribution systems, including both the shared portions and individual in-home "drop" wiring. As will be seen, the performance of such networks is subject to certain unavoidable limitations that become more severe as bandwidth is increased. As a result, modern systems generally use fiber-optic "supertrunks" in place of the long, repeatered coaxial trunks to interconnect small, physically separate coaxial distribution networks with shared headends. Chapter 12 and 13 will cover the technology and performance of such links. Chapter 14 will then cover alternatives to fiber optics (primarily microwave). Chapter 15 will cover the end-to-end downstream performance of the broadband network, from headend to subscriber, and Chapter 16 will deal with upstream issues.

Endnotes

1. *Reference Data for Radio Engineers*, 6th ed. Indianapolis, IN: Howard W. Sams, 1977, p. 6–6.
2. Frederick Terman, *Electronic and Radio Engineering*. New York: McGraw-Hill, 1955, p. 121.
3. *NCTA Recommended Practices for Measurements on Cable Television Systems*, 2nd ed. Washington, DC: NCTA, October 1993, Part Six.
4. The size 59, 6, and 11 cables were formerly known as RG59, RG6, and RG11 because of their physical size similarity to standard military-grade 75-ohm cables carrying those designations. Since CATV cables do not attempt to

adhere to the military specifications, the RG notation has been dropped by the industry.

5. The SCTE is an ANSI-accredited standards-making body. It publishes a wide variety of standards covering both performance and test methods related to cable television products. In particular, SCTE-IPS-TP-011, *Test Method for Transfer Impedance*, and SCTE-IPS-TP-403, *Test Method for Shielding Effectiveness*, detail test methods for evaluating shielding integrity, whereas SCTE-IPS-SP-001, *Specification for Flexible RF Coaxial Drop Cable*, contains both mechanical and electrical minimum performance standards.
6. *Code of Federal Regulations*: C.F.R. 47 §76.5(w). Technically, the noise is defined as that occupying the band between 1.25 and 5.25 MHz above the lower band edge. Available from the Government Printing Office, Washington, DC.
7. An excellent detailed mathematical treatment of amplifier distortion based on a power-series expansion is given in the *Technical Handbook for CATV Systems*, authored by Ken Simons and published by Jerrold Electronics Corporation, Hatboro, PA (the third edition was printed in 1968 and was still in print in 1978). His results were also published in the *Proceedings of the IEEE*, in July 1970.
8. Schlomo Ovadia, *Broadband Cable TV Access Networks*. Upper Saddle River, NJ: Prentice-Hall, 2001, pp. 49–50.
9. Data taken from *Cable TV Reference Guide*, April 1992 edition, Philips Broadband Networks, 100 Fairgrounds Drive, Manlius, NY 13104.
10. Private correspondence from Archer Taylor based on his experiences with installed systems.
11. This treatment of CID is based on the paper by Jeff Hamilton and Dean Stoneback, The Effect of Digital Carriers on Analog CATV Distribution Systems, *1993 NCTA Technical Papers*, pp. 100–113. NCTA, Washington, DC.
12. SCTE IPS-TP-201 *Test Method for Frequency Response and Gain*. SCTE, Exton, PA.
13. A basic discussion of feed-forward gain block principles is contained in the paper by John Pavlic, Some Considerations for Applying Several Feedforward Gain Block Models to CATV Distribution Amplifiers, *Technical Papers, Cable '83*. NCTA, Washington, DC., June 1983.
14. Michael G. Ellis, RF Directional Couplers. *RF Design*. February 1997, p. 33ff.
15. Robert P. Gilmore, Applications of Power Combining in Communications Systems. *Application Note*. Qualcomm, Inc., San Diego, CA.



Chapter 11
Coaxial Distribution System Design

11.1 Introduction

Distribution system design is, ideally, the process of specifying the most economical network that will provide the required bidirectional bandwidth to the required number of terminal points and still meet the performance goals. The performance goals, in turn, may be derived from internal quality standards, governmental regulations, the perceived requirements for reliable signal transportation, or all of the above.

Typical system specifications will include noise, distortion, response variation, delivered signal levels (and their stability), and hum. Reliability, another important consideration, is discussed in Chapter 20, and the effects of imperfect impedance matching among components are discussed in Chapter 15. The distribution system performance must also be considered as part of a chain that includes the headend channel processing and terminal equipment effects on signals. These are considered in separate chapters as well. Finally, the coaxial distribution system may well be cascaded with a signal transportation link such as microwave or, more frequently, a fiber-optic circuit. These, also, are treated in different chapters.

The final leg in the broadband distribution system is the drop wiring leading from tap ports to and throughout homes. By tradition, broadband networks are designed to meet certain goals from headend output to tap port, with an additional allowance made for the effects of in-home wiring. Modern drop systems, however, may vary widely in their complexity and quality, and in fact, may constitute minidistribution networks by themselves. The final section of this chapter deals with drops.

We begin by discussing basic principles and applying them to the "hard cable" portion of the plant.

11.2 Carrier-to-Noise Ratio

Noise is defined by the IEEE as "unwanted disturbances superposed upon a useful signal that tend to obscure its information content."[1] In many

communications systems, that is a good working definition. In cable television, however, the term *noise* is generally applied to thermal noise, which is treated independently from disturbances caused, for instance, by intermodulation products, ingress, or other causes. One of the reasons for treating thermal noise differently from other disturbances is that the subjective effects of noise and various other types of interference are different for analog video signals. Another is that, unlike many communications systems, the sources and mechanisms by which interfering signals are generated are well understood, and a detailed independent analysis of each has proved useful. Therefore, in this book, unless otherwise stated, *noise* will be used to refer exclusively to thermal noise.

In general, thermal noise is *uncorrelated*, that is, unpredictable and nonrepetitive. Noise generated in one device or at one time is totally unrelated to noise in another device or in the same device at another time although the *average* level of noise power may be quite predictable. Since that is the case, when the noise from two independent sources is combined, the total noise *power* will be the sum of the original noise powers.

Carrier-to-noise ratio (C/N) is defined as follows:

$$C/N(\text{dB}) \equiv 10\log\left(\frac{c}{n}\right) \tag{11.1}$$

where c and n are the scalar power levels of the carrier and noise, respectively. For analog television signals, such as NTSC, the carrier power is defined as the rms power that occurs during the sync tip (the highest power level in the modulated signal). Noise is defined as the rms noise power measured in a defined bandwidth (4 MHz for NTSC). Equation (10.19) gives the C/N for a channel transmitted through an individual amplifier, given the input signal level and noise figure.

In order to design a coaxial distribution system, it is necessary to compute the C/N of cascaded amplifiers, each followed by loss equal to the gain of the preceding amplifier. For two amplifiers, the general expression for composite C/N is approximated by the expression

$$C/N_{ttl} = -10\log\left[10^{-\left(\frac{C/N_1}{10}\right)} + 10^{-\left(\frac{C/N_2}{10}\right)}\right] \tag{11.2}$$

where C/N_1 and C/N_2 are the C/N of each of the amplifiers calculated independently using Equation (10.19) (for example, when driven by a noise-free input signal). Although this general expression is useful when cascading two dissimilar network elements, a simpler approach is useful when cascading n identical amplifiers, separated by passive network losses equal to the amplifier gain:

$$C/N_{\text{cascade}} = C/N_A - 10\log n \tag{11.3}$$

where

C/N_A = the C/N of a single amplifier

n = the number of similar amplifiers cascaded and separated by losses equal to their gains

These equations are quite useful, but their limitations must be emphasized.

- Equation (11.3) is applicable only to a cascade of amplifiers and loss sections where the loss of each section equals the gain of the preceding amplifier, and each amplifier has the same noise figure.
- Equations (11.2) and (11.3) are accurate only where the gain of each amplifier is high (a minimum gain of 20 dB is adequate to assure reasonable accuracy).
- Equations (11.2) and (11.3) are accurate only where the amplifier output noise power is high when compared with the thermal noise "floor" (a noise figure of 2 dB or greater combined with a 20-dB or greater gain is sufficient to assure reasonable accuracy).
- The input to the first amplifier contains only thermal noise.

The preceding conditions are required because the equations do not rigorously account for the effects of thermal noise though they are adequate for analyzing typical amplifier cascades encountered in cable television systems. For other applications, a more precise set of equations is recommended.

In a simple repeatered transmission line, the designer must "juggle" three parameters to reach a required end-of-line C/N: the noise figure of individual amplifiers, the input level to each amplifier, and the number of amplifiers to be cascaded. In more complex networks, the same variables apply, but the C/N "budget" may first have been allocated among several dissimilar network sections. As will be seen, noise considerations must be balanced against other requirements in reaching an optimal design.

11.3 Carrier to Distortion

Unlike the thermal noise, which is a function of only the transmission system characteristics and channel bandwidth, the distortion level depends on the signals carried. The distortion level of amplifiers is typically specified under one or two specific channel-loading conditions. Although it would be possible to simply specify distortion with just two or three carriers, the translation to expected IM products under the unique channel loading used in cable television is not straightforward, so vendors have resorted to testing under conditions that are easily related to actual field applications.

As discussed in Chapter 10, the channel loading under which amplifiers are typically tested and specified (for *CTB* and *CSO*) consists of independently generated, unmodulated carriers at each of the visual carrier frequencies specified in the ANSI/EIA-542 Standard frequency assignment list that falls within the bandwidth of the device. In addition, a portion of the bandwidth may be loaded with signals that simulate the effect of digitally modulated signals (often just band-limited white noise at a controlled level). Testing of XMOD is similar except that the carriers are simultaneously modulated.

The results of such testing do not match the distortion when loaded with analog television carriers (whose sync peak carrier levels match the unmodulated carriers), but the results are much more repeatable, and there are standard assumptions regarding the approximate improvement when loaded with modulated carriers: 6 dB in the case of *CSO* and 12 dB in the case of CTB and XMOD, though the improvement will be somewhat less with short coaxial cascades.

The buildup of distortion as amplifiers are cascaded is less straightforward than in the case of noise. If the deviations from linear response are not predictable from amplifier to amplifier, then the distortions can be considered uncorrelated, in which case they will add in the same way as noise; that is, the distortion products will increase by 10 log n, where n is the number of amplifiers. On the other hand, if every amplifier has the same general distortion shape (for instance, a flattening of the curve at voltage peaks), then the distortion will be highly correlated, and the magnitude of the products will increase at a 20 log n rate.

Since third-order distortion generally does result from symmetrical curve flattening and both *CTB* and XMOD arise from third-order distortion, most designs assume that these distortions increase at the 20 log n rate although that may be overly conservative when evaluating cascades of amplifiers with dissimilar characteristics or those that nominally cancel third-order distortion (for example, feed-forward amplifiers). Second-order distortions, on the other hand, are nominally canceled in the universally used push–pull gain modules. For that reason, the residual second-order distortions are not highly correlated, and many designs are based on a compromise 15 log n buildup.

As with cascades of identical amplifiers, the buildup of distortion when cascading different network sections depends on the degree to which the distortions are correlated.[2] In general,

$$C/dist = -x \log\left[10^{-\left(\frac{C/dist_1}{x}\right)} + 10^{-\left(\frac{C/dist_2}{x}\right)}\right] \tag{11.4}$$

where

$C/dist$ = the resultant carrier-to-distortion ratio in dB

x = from 10 to 20 depending on the degree of distortion correlation

$C/dist_1$ and $C/dist_2$ = the distortion ratios of the individual sections

Where identical amplifiers operating at the same output levels are cascaded, the expression for *C/CTB* can be reduced to

$$C/CTB_{\text{cascade}} = C/CTB_A - 20\log n \tag{11.5}$$

where

C/CTB_A = the distortion of a single amplifier
n = the number of cascaded amplifiers

whereas the equivalent expression for *CSO* is

$$C/CSO_{\text{cascade}} = C/CSO_A - x\log n \tag{11.6}$$

where

C/CSO_A = the distortion of a single amplifier
n = the number of cascaded amplifiers
x = from 10 to 20 depending on the degree of second-order distortion correlation

11.4 Noise–Distortion Trade-Off

The interplay between noise and distortion can be readily illustrated by considering a cascade of n identical amplifiers. It can readily be understood that the loss between each pair of amplifiers must be made identical to the gain of one amplifier for optimum performance. If the loss is less than that value, then each amplifier's input level (and, hence, output level) will be greater than the previous amplifier, and the distortions will quickly build up to a high level. If the loss is greater than the gain, then the input of each amplifier will be less than the previous amplifier and thus contribute disproportionately to the overall *C/N* degradation.

Given equal gains and losses, let us assume a requirement that the cascade provide an analog television *C/N* of 50 dB and a *C/CTB* of 65 dB. (We could have considered any other distortion parameter as easily, but this will serve to illustrate the trade-offs involved.) Let us also assume that each amplifier has a noise figure of 8 dB, a gain of 22 dB, and a *C/CTB* of 80 dB at the required channel loading and at an output level per carrier of +40 dBmV. (Again, to make it simple, we will assume all carriers have the same level and the loss between the amplifiers is the same at all frequencies.)

If our cascade consists of just a single amplifier ($n = 1$), then we can calculate the minimum input level that will assure an adequate *C/N*. From Equation (10.19):

$$C_i(\text{minimum}) = C/N\ (\text{dB}) - 59.2 + F\ (\text{dB}) = -1.2\ \text{dBmV} \tag{11.7}$$

With a station gain of 22 dB, this corresponds to an output level of +20.8 dBmV.

Similarly, since *CTB* changes 2 dB for every 1-dB change in operating level, we can readily determine that we could increase the output operating level to +47.5 dBmV and stay within the distortion limit. Thus, the output level could vary over a 26.7-dB range while maintaining the required performance.

If additional amplifiers are cascaded, the effective noise figure increases by 3 dB for every doubling of the cascade (Equation (11.3)), so the minimum C_i (and thus output level) must increase by that amount to provide the same end-of-line *C/N*. Similarly, the distortion increases by 6 dB for every doubling of the cascade, so the maximum allowable output level must decrease by 3 dB to maintain the same end-of-line distortion levels. Thus the allowable range of signal levels decreases by 6 dB.

Table 11.1 shows the allowable output level range as a function of cascade for this example.

If we plot these general relationships as a function of cascade, we can see that there is a maximum attainable cascade and a unique operating level that allows that cascade to be realized. Figure 11.1 illustrates the usable operating range, along with the parameters that define the noise–distortion–cascade relationship. Complete system design requires that this calculation be performed for both second- and third-order distortion, as well as noise, and to pick operating

Table 11.1 Example of cascade effects on allowable amplifier operating levels

	Allowable Output Level Range (dBmV) +	
Amplifier Cascade	*Minimum (C/N limited)*	*Maximum (CTB limited)*
1	+20.8	+47.5
2	+23.8	+44.5
4	+26.8	+41.5
8	+29.8	+38.5
16	+32.8	+35.5
21	+34.0	+34.3
> 21	Not possible and still meet both noise and distortion specs	

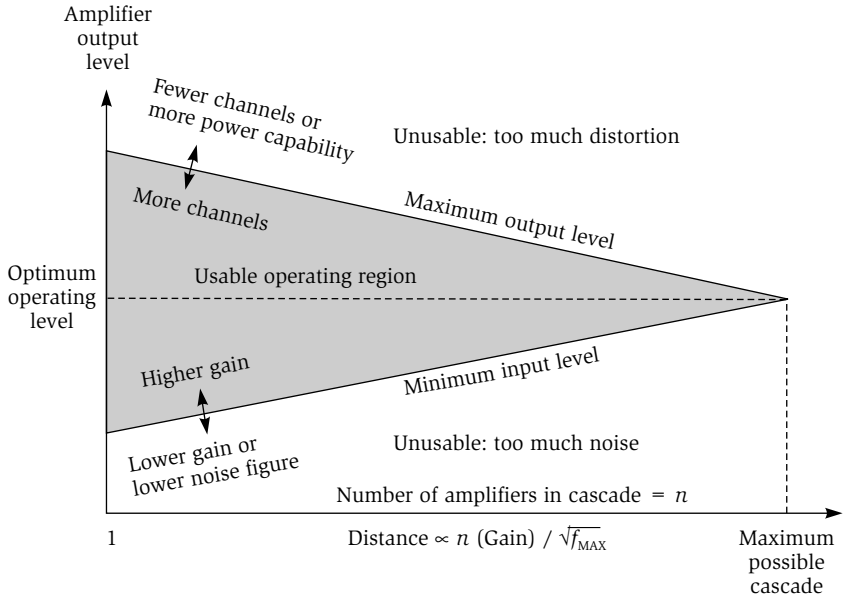


Figure 11.1 Relationship of cascade, noise, distortion, and levels.

levels that allow all specifications to be met. Generally, composite intermodulation noise (CIN) caused by digital signals is quantified as an equivalent increase in noise level that must be figured into the *C/N* calculation.

Another way of looking at the interaction between parameters is to consider the case of a hypothetical tapped distribution line that just meets end-of-line noise and distortion specifications, as well as required signal levels at subscriber terminal equipment. The operator desires to increase the bandwidth and channel loading of this system without degrading performance.

At the increased upper frequency limit, the cable loss is higher. The designer can compensate for that loss by using higher-gain amplifiers, spacing amplifiers closer together, or reducing the per-foot loss of the interconnecting cable. Additionally, the increased number of carried signals will increase the distortion products in each channel if the per-carrier level is held constant. Finally, the loss of subscriber drop wiring will increase, requiring higher tap levels or lower-loss cables.

If higher-gain amplifiers are used, then either output levels must be raised or input levels lowered. Absent a change in technology, the former will further increase distortion, whereas the latter will cause a greater *C/N* degradation in each amplifier station. If amplifiers are spaced more closely together, then more will be required to reach the same physical distance, thereby increasing both noise and distortion. Even if amplifier spacing is retained, either amplifier output

levels must be raised or drop cables replaced to maintain adequate levels at subscriber equipment at the highest channels. Absent changes in technology, the designer is faced with an insurmountable obstacle.

Several techniques can be used to get out of this "box": (1) amplifiers with lower noise figures and/or distortions can be used; (2) both distribution and drop cables can be replaced with larger, lower-loss versions; and/or (3) methods can be used to transport signals to points closer to homes so that coaxial cascades can be shorter. Generally, upgrades to older all-coaxial systems employ elements of all three techniques. In particular, fiber optics or microwave is used as a *supertrunk* to transport signals throughout large service areas with less noise and distortion than reasonably possible with coaxial techniques.

11.5 System Powering

As discussed in Chapter 10, coaxial distribution systems are usually powered by multiplexing 50- or 60-Hz power with the signals and transporting both through a common cable. Through use of separate signal paths through amplifier stations and passive devices, power routing is configured independently from the signal paths.

Coaxial distribution systems may be powered from one end (centralized powering) or via power supplies located along the cable route (distributed powering). In the latter case, it is common to locate the power supply midway through the section of plant it supplies so that roughly half the current can flow in each direction from the insertion point, reducing both the current capacity requirement for components and resistive losses in the interconnecting cables.

The reliability of distribution system power has a major effect on overall reliability. Therefore, systems designed for high availability are generally built with coaxial distribution networks that are centrally powered with high-capacity standby supplies. Chapter 20 covers reliability and availability.

Figure 11.2 is a simplified schematic of a section of cable and several amplifiers extending one direction from a power supply. From a network designer's point of view, the challenge is to minimize the number of power supplies required or, put another way, to determine the maximum number of amplifiers that can be powered from each supply.

Given a decision on system supply voltages, several parameters cannot be exceeded in making this determination: (1) the current rating of the supply, (2) the current rating of any coaxial component along the route, and (3) the minimum usable voltage at any powered component (typically around 40 volts for 60- or 90-volt systems).

As previously discussed, most modern amplifiers use switching power supplies. These draw current in approximately inverse proportion to the supply voltage so that the total power drawn is a constant. This allows them to maintain

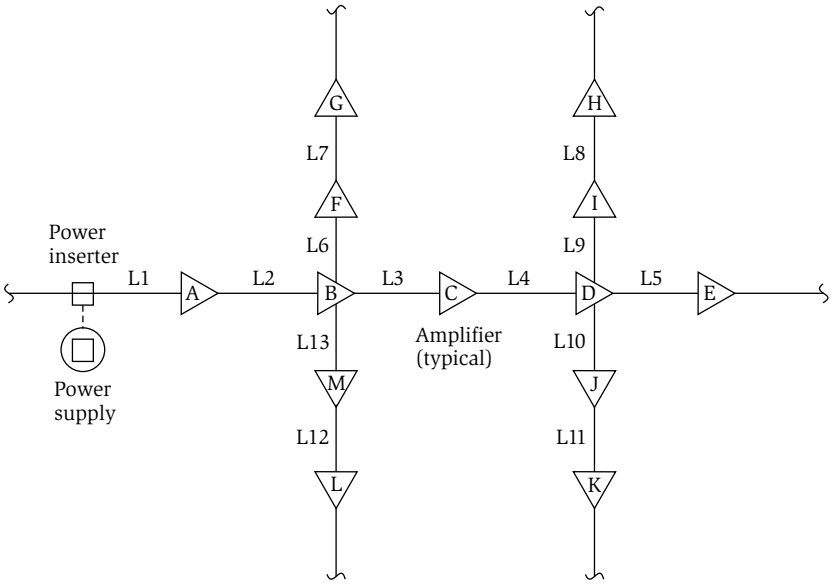
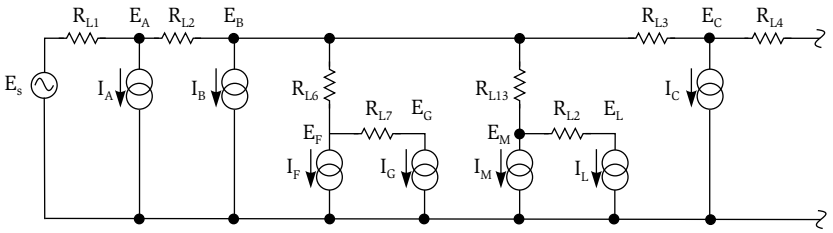


Figure 11.2 Example of amplifier powering.

high efficiency and low power dissipation. It also complicates the power calculation.

Figure 11.3 is a redrafting of the first seven amplifiers in Figure 11.2 to show the power component interaction. Though it would be possible to set up simultaneous equations for each powering situation and solve for all the critical voltages and currents, a successive approximation procedure is generally used



where $I_n = P_n/E_n$

P_n = power consumption of amplifier n

Figure 11.3 Equivalent circuit of powered coaxial network.

for both manual and computer calculations. The process is more or less as follows:

1. Determine the resistance of each cable leg at the maximum expected operating temperature.
2. Determine the voltage–current curve for each of the types of amplifiers to be used.
3. Determine the pass-through current ratings for each of the system components.
4. Make an initial estimate of the supply voltages at each of the amplifiers and the number of amplifiers to be served.
5. From that data, determine the current draw of each amplifier and, therefore, the total current drawn in each cable segment.
6. Using the cable currents from step 5 and the resistances from step 1, determine the voltage drop in each segment.
7. Starting at the power supply, subtract the first segment drop from the supply voltage to determine the voltage at amplifier 1. If different from the initial estimate, reestimate the amplifier voltages and currents and loop back to step 6.
8. When the iterative process results in adequately consistent results (very small adjustments needed), confirm that each of the current ratings has not been exceeded and that the voltage at the most distant amplifier is adequate. If not, drop one amplifier and start over.

The actual calculation is very easy on a computerized system. The "black magic" is in strategically placing power supplies and determining network power boundaries. Not only is it important to minimize the number of power supplies, but from a reliability standpoint, it is important to minimize the number of cascaded supplies that affect any given subscriber. Finally, it is important to locate power supplies where the commercial power is most reliable.[3]

A technique sometimes used to overcome both the maximum component current ratings and the voltage drop in the high-current sections near the power supplies is to add a second cable extending from power supply locations to additional strategically located power insertion points. Since these cables do not carry RF, they can be of lower characteristic impedance and, thus, use larger center conductors. For instance, the loop resistance of standard 0.625-inch-diameter P3 cable is 1.1 ohms/1,000 feet, whereas the same-size "power feeder" cable is 0.290 ohms/1,000 feet.

11.5.1 Switching Transient Effects

A consequence of the interaction among switching power supplies located along a coaxial line is that anything that causes one amplifier to draw additional

current will cause all other amplifiers along the same line to also draw additional current. This is because, regardless of where it is located, the additional current supplied to one amplifier will increase the current through the first coaxial segment, which will, in turn, result in lower supply voltages and therefore higher currents for all other amplifiers.

Though amplifiers normally draw a constant amount of power, consider what happens when a non-UPS system power supply switches between commercial and standby power. As discussed in Chapter 10, typical switching times are 8–16 ms. During this period, no power is supplied to the network.

When the power is restored, each of the amplifier power packs will attempt to draw additional current to make up for the missing 1–2 half cycles. The result is that the voltage at the most distant amplifier may well drop below operation range. In fact, one study has shown that interruptions as short as 8 ms can grow to effective interruptions of greater than 50 ms at the fifth amplifier away from a power supply.[4] It is incumbent upon network designers to evaluate whether such interruptions are unacceptable for anticipated services and signal formats and to consider UPS switching if appropriate.

11.5.2 Hum Buildup

The mechanism for hum modulation is similar in passives and the signal-processing components of amplifier stations. Since the period of the power signal is much longer than the transit time through the network (the wavelength of 60 Hz is about 3,000 miles), a conservative assumption would be that effects of the multiplexed power will be in phase throughout the system and that, therefore, hum modulation due to parametric effects should build up in a 20 log fashion. As a practical matter, however, much high-voltage utility distribution is three phase, and as a consequence, power supplies throughout a large service area are not likely to all be connected to the same power phase.

To illustrate the worst-case magnitude, assume that the cascade in a distribution line includes 5 amplifier stations, 4 splitters and/or couplers, and 30 taps. If each has a hum rating of -70 dB and all components are subjected to maximum voltage and current stress, the end-of-line hum would be $-70 + 20\log(5 + 4 + 30) = -38$ dB, equivalent to 1.2%. Powering, however, will likely be distributed across several phases, and most components, especially end-of-line taps, are unlikely to experience anywhere close to their rated voltage and current so that total hum buildup should be well under 1% if all components are operating properly.

11.6 Signal Level Management

The requirements for level management are very different in the forward and reverse directions. In the downstream or forward direction, each amplifier

receives signals from only a single source whose output is well defined. The design and operational requirements are constant amplifier output levels and adequate amplifier input levels. Additionally, to ensure that well-designed end-of-line receivers work satisfactorily, several parameters must be controlled: adequate signal strength to overcome receiver noise, avoidance of excess levels that could lead to overload, limits on adjacent channel level differences, limits on total instantaneous variation across the spectrum, and limits on changes with respect to time. These are more fully discussed in Chapter 24.

As discussed in Chapter 10, manufacturers typically specify performance with specific channel loadings and amplitudes. Since digital formats are generally more robust than analog television, it is common to specify performance under two conditions: a full spectrum of analog television signals and a defined mix of analog television and digital signals, where the digital signals are operated at a relatively lower level (as illustrated in Figure 10.19). Note that the *net gain* of the network, as measured from the output of one amplifier to the output of the next amplifier, is 0 dB in either case. The depressed levels of the digital signals relative to the analog signals are set in the headend where they are combined.

To the extent that the positive gain slope built into amplifiers does not exactly compensate for the losses between amplifiers, *equalizers* are used to eliminate the remaining variability. To the extent that interamplifier losses (including any equalizer losses) are less than amplifier gain, attenuators (also known as *pads*) are used to set the gain block input level. Pads may be placed before the first stage (generally, the only option with line extenders) and sometimes also between stages. If the latter option is available, then in cases where excess signal level is available, operators can independently set the operating levels of input and subsequent gain blocks to optimize noise–distortion trade-offs. With amplifiers that offer multiple independent output stages, the interstage pads also allow setting of different output levels for different ports. Thus, a port that directly feeds a tapped line may be set to a higher level than one that will feed additional amplifiers and therefore needs to run at a lower distortion level.

In tapped lines, both distribution cable losses and drop cable losses must be considered. It is important that adequate and relatively uniform levels be delivered to terminal equipment (see Chapter 24). A 200-foot size 6 drop cable has about 8 dB greater loss at 750 MHz than at 54 MHz. This differential loss tends to offset some of the intentional tilt in amplifier outputs if connected to a tap that is close to a station. On the other hand, if it is connected to the last tap in a string, it will aggravate the situation if the slope is already negative (high-end channels below the low-end channels in level). To prevent excessive level variation, in-line equalizers are sometimes inserted in tap strings. As a general rule, operators try to limit analog video channel level variation across the spectrum to about 10 dB. Typical design rules call for +15 to +20 dBmV levels at tap ports on the highest channels.

Signal level management in the upstream direction is much more complex and must be treated as an overall system problem rather than as just part of the design. Signals may originate from anywhere in the system and may be of various types so that the channel loading and levels may vary continuously at any given amplifier.

The preferred approach to upstream level and gain setting is to align the system from headend out so that the *input* level to each upstream amplifier is consistent; that is, if a level of x dBmV is inserted into the input of any upstream amplifier, the level received at the headend will be the same. Pads (and, if required, equalizers) are placed in the *output* of the gain modules to set overall system gains properly.

This does not, however, mean that the required upstream transmitter output levels for terminal equipment will be consistent. Consider Figure 11.4, which depicts a typical tapped line. Following normal practice, tap values are selected to produce approximately the same downstream level at the highest channel at each home. The circuit losses depicted are based on typical manufacturers' specifications for four-port taps, the loss of 250 feet of .500-inch type P3 cable between the taps, and the loss of 100 feet of size 6 drop cable connected to each tap port. As can be seen, the downstream levels at the end of the drops are all within 1 dB.

If we calculate the insertion levels required to reach the return amplifier, however, we see a nearly 20-dB difference due to the differential loss of the cable and the taps. Rather than manually set the level of each upstream transmitter, most manufacturers sense the level reaching the headend and remotely control each device independently to set the appropriate level. Another approach is to

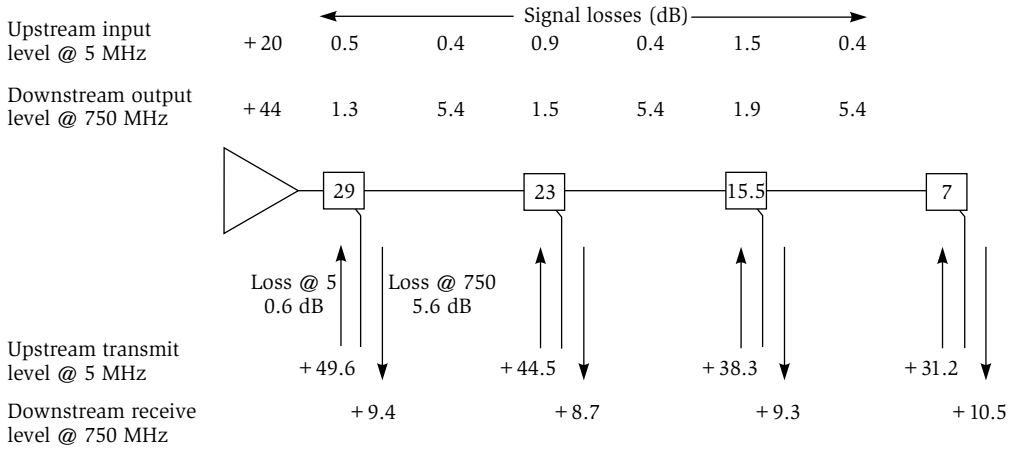


Figure 11.4 Signal levels in tapped distribution line.

equalize each drop to approximately compensate for the differential cable loss and equalize required levels. As of the writing of this book, there is no uniformly accepted practice in this area.[5] Chapter 16 deals with the numerous unique issues related to upstream transmission.

11.7 Signal Level Stability

As was discussed in Chapter 10, the attenuation of coaxial cable varies about 1% per 10°F. The losses of other components, as well as amplifier gains, also vary somewhat with time and temperature. Amplifier manufacturers offer two options in their equipment that are designed to compensate these variations: thermal compensators and automatic gain and/or slope control circuits (AGC/ASC).

Thermal gain compensation is *open loop*, meaning that the gain adjustment is only an approximation to the variation in interconnecting losses with temperature. AGC circuits work by sensing the level of one or more of the carried signals and controlling the gain of the station so as to maintain a constant output level for that signal. Automatic slope controls sense two signal amplitudes and control both the overall gain and the gain slope through the stage so that the levels of both signals are constant at the output port. Typical control ranges are 6–8 dB. Use of AGC and/or ASC circuits depends on the accuracy of headend level setting. Also, if one of the reference carriers fails, the entire distribution system can change gain sufficiently to cause serious distortion products.

The loss between amplifiers may be all cable or a combination of cable and passive devices. In the extreme case of an all-cable link whose loss is 35 dB, the loss will vary over a range of about 5.2 dB over a range from +110°F to −40°F(+ 43°C to − 40°C). Thus every amplifier will require some sort of compensation in a trunk network.

In a tapped distribution line or one that is split several times, half or more of the loss may take place in passive devices. Thus, thermal compensation may be required only in every second or third amplifier. The trade-off is that compensation networks may have loss that reduces available gain, whereas level variations will affect *C/N*, *CTB*, and *CSO*.

11.8 The Service Drop

So far we have dealt primarily with the portion of the distribution network that lies between the headend output and the tap ports that serve individual customers. Generally, network designers design to specifications for this portion of the network, assuming certain average performance in the headend and drop systems such that the end-to-end performance will meet regulatory requirements as well as subscriber expectations. Chapter 2 dealt with subscriber quality

expectations, and Chapter 15 will cover end-to-end performance. This section will deal with the variation in drop subsystems and suggest the performance degradation that might be expected in a "standard" drop that must be taken into account in calculating end-to-end performance.

As a practical matter, in the United States, as in many other countries, all or most of the drop system may not be designed, built, or maintained by the network operator. Thus, although designers must be aware of the variation in drop performance, little can be done to control quality or design within homes.

11.8.1 Simple Passive Architectures

At its simplest, the drop will consist of a flexible (either overhead or underground) cable from the distribution tap port to the building entry point, a connection to the building ground system at that point, and another flexible cable within the building that leads to a user's receiver, such as a television set or VCR. If there are multiple receiver locations, a signal splitter is used to divide the signal to feed more than one outlet.

If the loss in the interconnecting cables plus splitters is too great to assure adequate signal level at receiver input ports, then either (1) higher tap levels must be provided, (2) lower-loss drop cable must be used, or (3) amplification must be provided in the drop system. As discussed earlier in this chapter, tap signal levels generally are difficult to increase because of the interaction among levels, noise, and distortion in the distribution network. That leaves the builder of the drop with two practical choices: reduce loss or provide gain.

In the early history of cable systems, almost all drop wiring was done with size 59 cable (originally known as RG59). As bandwidths have increased beyond 300 MHz, common practice has been to upgrade the portion of the drop between tap and bonding (ground) block (the point where the cable enters the dwelling) to at least size 6 and, as required, size 7 and even size 11. The shift from 59 to 6 is easily justified by savings in the hard cable plant due to lower required tap levels. The larger sizes, however, are considerably more expensive and, in the case of size 11, require special connectors that add to both hardware cost and installation labor. Even in systems with size 6 outside drop cables, the inside wiring is generally size 59 since it is easier to handle and less obtrusive visually where exposed. Common practice is to design the hard cable plant to have drop levels that will accommodate most situations with size 6 exterior and size 59 interior cables, and to use larger cable sizes, both inside and outside, only where required because of exceptional drop lengths and/or a high number of splits.

A Reference Drop Structure

Although there are many single-receiver households, it is increasingly common to find two (or more) television outlets in cabled homes, and therefore a splitter

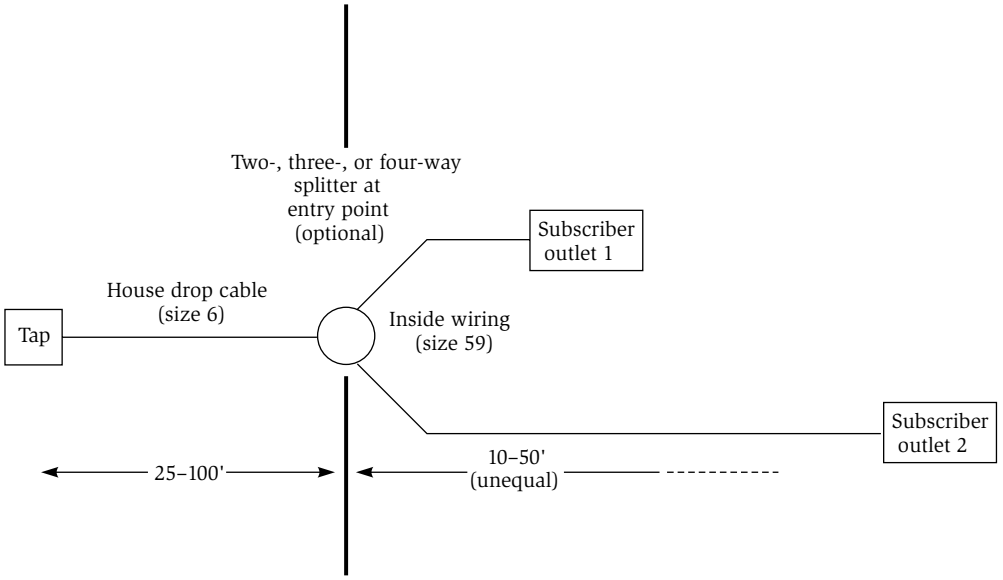


Figure 11.5 Typical drop configuration.

that may be mounted at the entry point to the dwelling and does double duty as a bonding block. Given that, we can define a *standard drop* (Figure 11.5) that will encompass the majority of current home wiring. It will include 25–100 feet of size 6 cable from tap to house, perhaps a splitter, and size 59 connections to outlets, with lengths varying from 10 to 50 feet.

Passive Drop Transmission Characteristics

In order to assess the transmission characteristics of this standard drop, we need to first predict the quality of the components used, then use the tools developed in Chapter 10 to calculate their effect on the signals.

Component Quality. A prime determinant of the degree of degradation signals experience in the drop system is the quality of the components used. Unfortunately, just as there are no mandatory standards on drop wiring or design, so there are none for the components used. Historically, drop wiring was installed and maintained by cable operators (much of it still is), and they purchased reasonable quality components because it was less expensive than the trouble calls and customer dissatisfaction resulting from inferior materials.

As of the writing of this book, the Society of Cable Telecommunications Engineers (SCTE, an ANSI standards-making organization) is actively devel-

oping voluntary national standards covering virtually all drop components: cable, connectors, splitters, couplers, amplifiers, attenuators, filters, and so on.[6] Covered performance parameters include return losses, transmission losses, environmental requirements, shielding, surge protection, labeling, and a host of others. The specifications for amplifiers additionally include gain, noise, and distortion. Of importance in estimating drop performance, most devices must exhibit return losses of at least 18 dB, and excess loss (that is, loss exceeding that expected from the power division ratio) in splitters and couplers may not exceed 1.0 to 1.2 dB (for instance, the loss through a two-way splitter must be no greater than 4.0 dB) though in some cases, they can rise above 2 dB. Port-to-port isolation in splitters must be at least 20 dB. Cable-shielding requirements are consistent with good-quality double-shielded cables. There is no assurance, of course, that existing products all meet the proposed standards or that even conforming products will continue to do so after years of exposure to weather.

Furthermore, network operators can hardly be assured that components installed by a homeowner will meet these voluntary, future standards. In particular, a conservative estimate is that return losses from components may be as poor at 10 dB, port-to-port splitter isolation may be as low as 10 dB, and cable losses may, after aging if not originally, exceed nominal values (Figure 10.5) by 10%.

Drop Cable. The characteristics of various types of drop cable were discussed in Chapter 10. Within the basic size options, however, are many quality and feature options. Chief among those are shielding integrity. Cables are available with a standard dual shield (foil and braid), triple shield (foil, braid, and foil) and quad shield (foil, braid, foil, and braid), with increasing effectiveness of shielding. Of serious concern to operators are dual shield (or worse) cables with sparse braid structures that are available in the retail market. Some of these poorly shielded cables do not adhere to SCTE tolerances for dimensions either, with the result that connectors are poorly attached, both electrically and mechanically, to the cable.

Aside from shielding options, cables are available with support messengers,with flooding compound (to prevent water from migrating through the cable structure), in a dual configuration (for dual cable distribution networks), and paired with one or more twisted copper pairs for carrying telephony or power along with RF in the same structure.

Finally, the nonmetallic portions of the cable are available with various degrees of flame resistance to meet electrical and fire codes when they are routed within some interior paths (such as plenums).

F Connectors. Whatever other components may be included in the drop, in North American cable systems, there will be at least four type F connectors — one at the tap, two at the bonding block, and one at the consumer's receiver

input. This connector, designed for low-cost mass manufacturing, is at the root of many drop problems. The female connector uses a nonprecious-metal, two-point spring clip contact for the center conductor; the male cable connector is a *feed-through* type that utilizes the center conductor of the cable itself for a pin; and the mated pair depends on a threaded joint plus a butt contact for an outer RF connection. The result is a connector family that is extremely craft sensitive in installation. Despite ANSI standardization,[7] and detailed electrochemical studies,[8] F connectors remain the leading cause of trouble calls and signal leakage in many cable systems.

In cases where power is multiplexed with RF signals on drop cables (typically to power point-of-entry telephone or data terminal equipment), the limitations of F connectors are exacerbated. A less-than-perfect contact that has a reasonable capacitance may pass RF signals but will not pass power. Furthermore, a loose contact, whether in the center conductor or shield, may generate electrical arcing when cables move with temperature changes or wind. These transients generally will have frequency components extending through the upstream band and into the lower downstream channels. Since they occur *within* the cable, they are tightly coupled to the transmission line in both directions and can cause severe transmission problems. The effects of electrical transients on upstream transmission are discussed in detail in Chapter 16.

In addition to inherent problems with the quality of the mating surfaces, male connectors vary in how they are attached to the flexible drop cable. Until about 1990, the most common construction used a sleeve that was pushed between the foil and braid of the cable. It was retained by use of a separate hexagonal *crimp ring*. Improved designs used a heavier, integral crimp ring. With both designs, however, field experience showed that the retention strength of the connector on the cable reduced over time. More recent designs have featured a full annular press-fit connection that has proved to retain its strength.

The use of a feed-through design (where the cable center conductor forms the connector pin) with more than one size of cable has created problems with the mating connectors. Even with highly compliant female contact structures, once a female connector has been mated with a large cable center conductor, its ability to form a reliable contact with a small cable is reduced. The center conductor of size 11 cable is beyond the capacity of standard female ports so that special pin-type male connectors are used with it. Pin connectors are also sometimes used with size 7 cable to protect mating ports.

Traps. In addition to the normal drop components, some combination of band-pass, band-stop, low-pass, or high-pass filters are often inserted into the drop circuit at the tap to control which signals reach subscribers' receivers. The techniques used to restrict and tailor services are discussed in Chapter 18. However they are used, traps add loss. Generally, their out-of-band loss may

be only a few tenths of a decibel, but their loss near cutoff frequencies can be as much as several decibels at the band edge of adjacent channels, coupled with a considerable slope through the channel. Where traps are used to control several channels individually, as many as four or five may be cascaded between tap and drop cable.

Transmission Loss. Given these variables, the total loss from tap to house will generally fall within the values given in Table 11.2, where the minimum losses result from short drop cables without a splitter, and the maximum losses are based on long drops, splitters, and two to three in-line traps.

System designers must take into account this loss range in setting signal levels, and allowable level variation, at system taps. For example, in the "maximum" loss configuration, note that the *difference* in loss is 7 dB over the downstream bandwidth of a 750-MHz system. If the flatness of the signal at the tap port has a downward slope with frequency, the combination could result in widely differing signal levels being presented to receiver input ports.

Microreflections, Group Delay, and Frequency Response Variation. Characteristics of terminal equipment are dealt with in depth in Chapters 23 and 24. For now, it is sufficient to state that cable operators are required to deliver signals directly to consumers' television receivers under some circumstances (where analog channels are unscrambled and delivered on standard channels) and that the return loss of those receivers is essentially zero on untuned channels and may be as poor as 4–8 dB on the tuned channel. In any event, it is likely that an unused outlet will be left unterminated, so it should be assumed that 100% reflection occurs at each subscriber outlet. This will cause reflections whose delay and magnitude will depend on the isolation of the splitter ports and the length of the cables beyond the splitter. Using the methods discussed in Chapter

Table 11.2 Tap to receiver loss ranges for typical passive drop systems

	Expected Loss Range	
Frequency	*Minimum*	*Maximum*
5 MHz	0.2 dB	9.5 dB
40 MHz	0.5 dB	10.5 dB
54 MHz	0.6 dB	13 dB
550 MHz	1.8 dB	18.5 dB
750 MHz	2.1 dB	20 dB

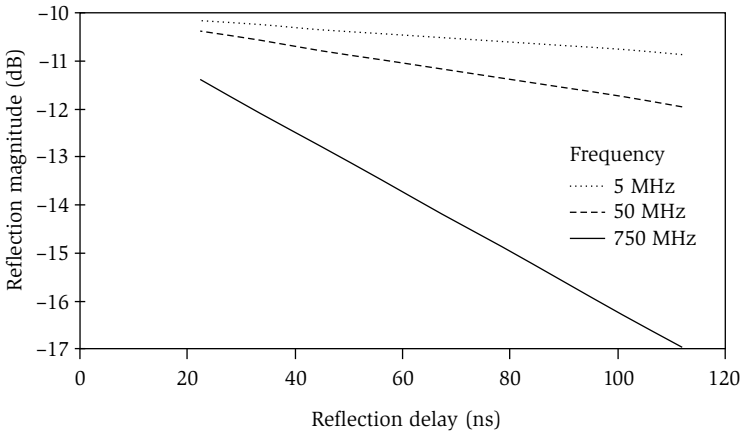


Figure 11.6 Worst-case microreflection magnitudes and delays in typical house wiring.

15, we can estimate the worst-case reflections. Figure 11.6 shows the results at various frequencies. Although operators should be prepared for occasional reflections of those magnitudes, a statistical study of actual cable systems showed 95% of all echoes, including those in the distribution plant, were more than 25 dB below carrier levels in magnitude.[9]

These reflections can lead to bit errors in digital transmissions since delayed signals cause successive transmitted symbols to overlap at the receiver and visible degradation of analog television pictures. As will be discussed in Chapter 15, the effects also include group delay and passband response variations. Amplitude response variations exceeding 1 dB in each channel are quite possible.

Signal Leakage and Ingress. Most difficult to assess is signal leakage. In practice, most leakage and ingress is not associated with components or cable but rather is due to inadequate shielding in customer-owned terminal equipment and/or improperly installed or tightened drop connectors. Poor-quality connectors and poorly shielded cables may also contribute problems.

Given the statistical nature of both egress and ingress, there is no way of predicting their effects. The effect of shielding of consumer equipment is discussed in Chapter 24. The effect of ingress on upstream communications is treated, along with other upstream transmission issues, in Chapter 16.

11.8.2 Adding Amplification—The Trade-Offs

The alternative of providing gain is attractive when it is less expensive than drop cable replacement or where even the largest cables would not be

adequate. However, many corollary issues are raised when amplification is added, including:

- Additional noise and distortion
- Signal reliability decrease owing to the amplifier failure rate combined with the availability of its powering source
- Gain stability and flatness, hum modulation, and all the other issues associated with active equipment
- Provision of a mounting location that has adequate environmental protection and available power, and yet is near where the cables to each of the outlets come together

If amplification is included in the drop system, the hard cable plant must perform at a higher performance level in virtually all respects in order to assure end-of-line quality. As a practical choice, drop amplifiers are seldom used except in multiple dwellings or commercial installations where they are, in fact, part of the distribution plant. One exception is in isolated single dwellings where there is no better solution to low levels.

Another even more rare instance is where the reverse isolation of the amplifier is needed to protect the system from ingressing signals that would otherwise travel up the drop and interfere with others' reception. This latter situation is discussed in depth in Chapter 19. Finally, there are homes where the need is not just to deliver cable signals to receivers but to interconnect multiple equipment.

11.8.3 Integrated In-Home Wiring Systems

In response to a need to have defined performance levels and a desire to interconnect home video equipment and multiple program sources, several industry groups have designed integrated home wiring systems. Although a complete discussion of these is beyond the scope of this book, one example of such a system is the CEBus.

The Consumer Electronics Manufacturers Association (CEMA) of the Electronic Industries Association (EIA) has developed a complex home wiring standard under the umbrella title CEBus. CEBus supports message and information transmission over a variety of physical paths within homes, each with its own special label: twisted pair wiring (TPBus), subcarriers on the power wiring (PLBus), coaxial cable (CXBus), infrared through the air (IRBus), over-the-air transmitted low-power RF (RFBus). A separate, but related, specification covers direct interconnection of home audio and video equipment using specially designed multiple twisted pair cables (AVBus). These standards have been under development since the mid-1980s.[10]

A common language is specified for medium-independent signaling of messages and control information, whereas routers interconnect different types of

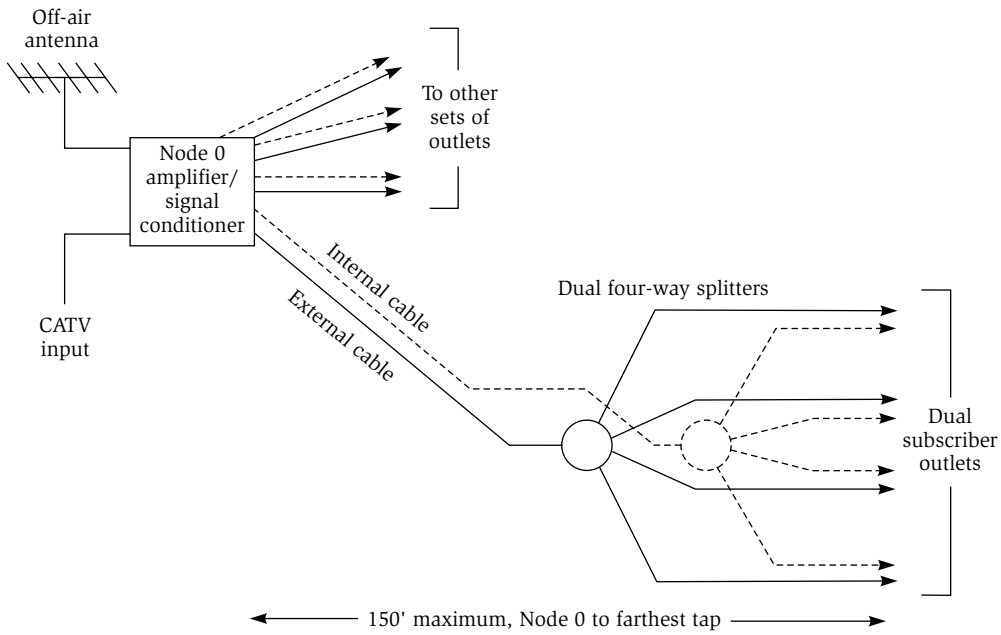


Figure 11.7 Basic CEBus coaxial topology.

media. The language used for signaling is known as the *common application language* (CAL) and is defined by ANSI/EIA Standard IS-60.

Of greatest interest to broadband network operators, the CXBus distributes both externally supplied signals (such as from a cable system) and internally generated RF signals over dual, size 6 drop cables through the residence. Downstream signals from the external network are amplified, optionally combined with internal signals, then split to feed four or more "branch circuits" over one of the paired distribution cables. Each branch cable may feed a four-way splitter with shorter cables from there to individual termination points. Thus, one network may incorporate as many outlets as permitted by the cable losses and available RF power (see Figure 11.7). Note that the splitters may be internal to Node 0 and that this is, in fact, the preferred topology since all network elements and connectors are then accessible for maintenance or inspection, if required.

The downstream bandwidth of the distribution system (the "external" cable in CEBus parlance) extends from 54 to more than 800 MHz, so as to allow direct transmission of UHF signals from local antennas, if desired (switching between the external network and local antennas is provided in the design). Upstream signals are carried in the 5–30-MHz band and are combined and amplified before insertion into the external network. The bandwidth from 450 to 546 MHz may

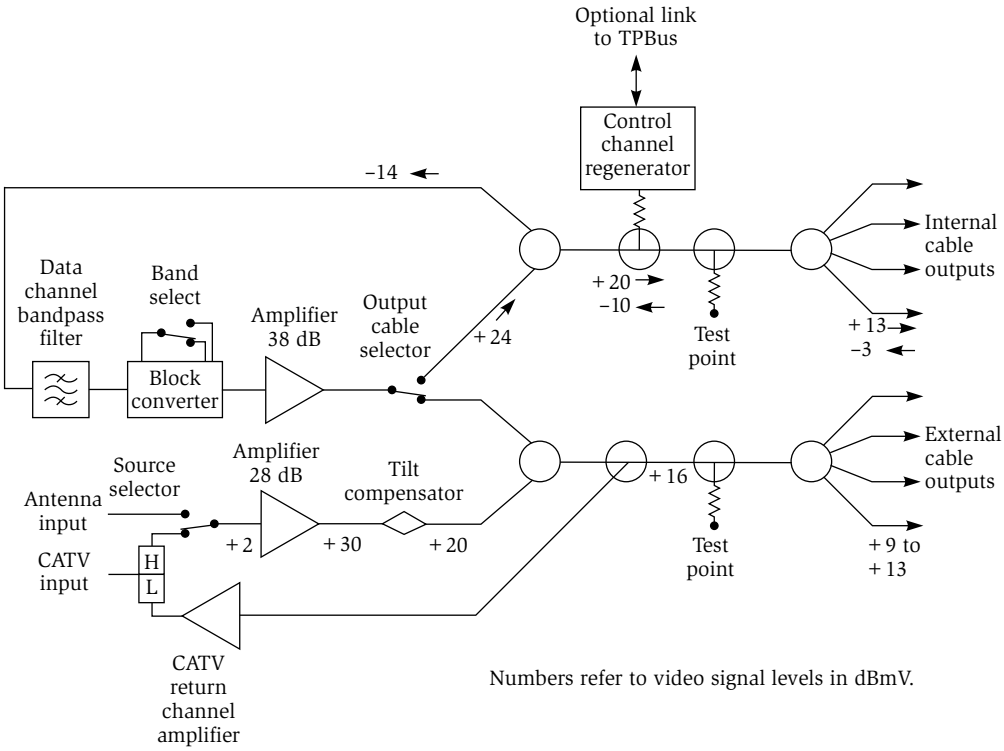


Figure 11.8 CEBus Node 0 functionality.

optionally be used for downstream transmission of video from local sources to television receivers. The second ("internal") cable has a more evenly split bandwidth and is used for routing of both messages and video among various consumer electronics devices. Data (information) channels are defined in 1.5-MHz bandwidth slices, with four contiguous channels required for internal transmission of normal analog video. Physically, the two cables are run together, with dual type F wall outlets provided for subscriber access.

All the signal processing is done at "Node 0." Figure 11.8 shows the signal processing within the node. Locally generated signals that will stay within the dwelling may be inserted into any internal cable outlet. The block converter is used to convert the frequencies of the internally generated signals from lower upstream channels to downstream channels for redistribution. The output cable selector determines whether the converted channels are distributed over the internal or external cables.

Locally generated signals that are intended for transmission through the external cable system can be inserted into any external cable outlet. Those

signals will be amplified by the return amplifier and transmitted into the cable drop.

The control channel regenerator is an optional feature that allows control signals to travel both directions between coaxial and twisted pair distribution networks.

The CXBus is a well-designed specification that should assure proper level control, minimal signal leakage or ingress, and well-controlled noise and distortion degradation. Nevertheless, there are important issues when external network signals are transmitted over such a network:

- The network has fixed bandwidth boundaries, both downstream and upstream. If a network operator chooses to use frequencies outside those bandwidths, the signals will not reach subscriber outlets.
- As we discussed, the downstream and upstream amplifiers contribute both noise and distortion that must be accounted for in the end-to-end performance budget. Since only a few homes may be equipped with CEBus systems, it is unrealistic to upgrade the entire external network to accommodate them.
- Upstream levels may vary widely in power level (see Chapter 16). Therefore, the return amplifier must have a very wide dynamic range.

Despite these concerns, it is expected that CEBus systems will generally interface with most current cable television network with minimal problems.

11.9 Summary

Cable television coaxial distribution systems consist of a cascade of cable and amplifiers. The network typically branches in the downstream direction, and various passive RF devices are used both within and external to amplifiers to create the legs. The amplifiers are capable of amplifying signals flowing in both directions, with the downstream bandwidth extending from approximately the lowest broadcast television channel (54 MHz in the United States) to an upper limit determined by the bandwidth needs of the network, but typically 400–870 MHz. The upstream bandwidth typically extends from 5 MHz to somewhere between 30 and 42 MHz.

In the downstream direction, distribution networks with more than a few amplifiers in cascade are usually designed and operated so that the gain, measured from amplifier output to amplifier output, is unity. When systems carry only one signal type (for example, analog video signals), their levels are often adjusted so that they increase linearly with frequency, as measured at amplifier output ports. This usually results in amplifier input levels that are approximately the same across the spectrum and creates the optimum balance

between noise and distortion. When different signal formats are carried, levels may be adjusted to a different pattern of levels. A common example is digital signals carrying packetized data or video, which are operated 6–10 dB lower in level than equivalent analog video signals would be at the same frequencies.

Amplifier gains must be sufficient to offset the loss of the intervening cable and passive devices at the highest operating frequency. A combination of attenuators, positive gain slope with frequency, and special response-shaping networks are used to just offset interamplifier losses. The last few amplifiers may be operated at higher levels to allow feeding more subscriber taps between gain blocks.

A distribution network will exhibit an overall downstream carrier-to-noise ratio that is inversely proportional to the logarithm of the number of amplifiers and proportional to the signal levels at the inputs to each amplifier. It will generate distortion products that are primarily determined by the amplifier output levels, as well as the number of amplifiers. The bandwidth, operating levels, noise, and distortion are interdependent and, as a set, describe the essential features of the system. Other important characteristics include hum modulation, reflections, frequency response, and gain stability.

Power is carried to amplifiers over the same cables used for signals. Most commonly, 60–90 VAC, 60-Hz quasi sine waves are used in the United States. In the event of commercial power failure, standby power supplies replace the sine waves with trapezoidal waves. The saturation of magnetic circuits used to separate 60-Hz power from signals at each device is the largest contributor to hum modulation.

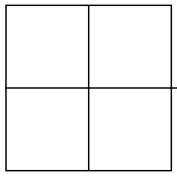
Cable television drop systems can be as simple as a single coaxial cable from tap to one television receiver, or as complex as a full, integrated distribution system with multiple video sources and receivers. Today, the vast majority of cable drop configurations consist of a cable from tap to home, an optional splitter, and additional cables to one to four individual outlets.

The loss (and its variation) over this network, along with reflections from the termination points and shielding integrity, will have an effect on the overall transmission quality to consumers' receivers. In more complex networks that include amplification, the noise, distortion, gain stability, group delay, hum modulation, and so on must be included in end-to-end performance calculations. In the case of complex distribution systems, spectral usage by internal video sources may affect compatibility with the external network.

Endnotes

1. *IEEE Standard Dictionary of Electrical and Electronics Terms*, 2nd ed. IEEE Std 100–1977, Institute of Electrical and Electronic Engineers, Inc., New York.

2. Several excellent papers have been written on this subject, including Doug McEwen et al., Distortion Characteristics of Integrated Fibre, AML and Cable Amplifier Systems, published in the *CCTA Convention Papers, 1990*, Canadian Cable Television Association, Suite 400, 85 Rue Albert Street, Ottawa, Ontario, K1P 6A4, Canada; and Doug McEwen et al., Distortion Accumulation in Trunks, published in the *1990 NCTA Technical Papers*, NCTA, Washington, DC.
3. An excellent treatment of this subject, along with practical recommendations, is contained in CableLabs' 1992 *Outage Reduction*, in the section entitled Power Grid Interconnection Optimization. CableLabs, Louisville, CO.
4. Doug Welch, Cable Powering into a Distributed Load. Presented at the 1995 NCTA National Convention and published in the *1995 NCTA Technical Papers*. NCTA, Washington, DC.
5. Dean A. Stoneback and William F. Beck, Designing the Return System for Full Digital Services, *Proceedings Manual, 1996 Conference on Emerging Technologies*. SCTE, Exton, PA, January 1996.
6. Society of Cable Telecommunications Engineers, Engineering Committee, Interface Practices Subcommittee. For current specifications and associated test procedures, contact the SCTE at 140 Philips Road, Exton, PA 19341.
7. ANSI/SCTE SP-400 specifies the mechanical dimensions of the female F port, and SCTE-IPS-SP-407 is the parent specification. Male cable F connectors are covered by SCTE-IPS-SP-408, and their dimensions are covered by SCTE-IPS-SP-401. SCTE, Exton, PA.
8. Brian Bauer, F-connector Corrosion in Aggressive Environments — An Electrochemical and Practical Evaluation. *1991 NCTA Technical Papers*, NCTA; and Brian Bauer, Hidden Influences on Drop Reliability: Effects of Low-Level Currents on F-Interface Corrosion and Performance. *1992 NCTA Technical Papers*, NCTA, Washington, DC.
9. Richard Prodan et al., Analysis of Cable System Digital Transmission Characteristics, *1994 NCTA Technical Papers*. NCTA, Washington, DC, May 1994, pp. 254–262.
10. For current information on the CEBus standard, contact the Electronic Industries Association, 2500 Wilson Blvd., Arlington, VA 22201.



Chapter 12

Linear Fiber-Optic Signal Transportation

12.1 Introduction

Although coaxial cable is used for the "last mile" delivery of broadband services to subscribers, other technologies are frequently used to cover most of the physical distance between headend and tap. One reason for this is that the interrelationship between bandwidth, signal levels, number of cascaded devices, noise, and distortion (discussed in Chapter 11) limits the size of purely coaxial systems. Second, large coaxial systems are best suited to delivery of a common spectrum of signals to all parts of the network, whereas many services, such as data communications and telephony, require a bidirectional bandwidth allocation to every customer. In order to provide this capacity in a network of attainable bandwidth, the network must be segmented, and this requires the use of small coaxial sections, each connected by separate communications links to the headend.

In modern networks, these links almost always use linear fiber optics for at least the segment that interfaces with the coaxial network (though Chapter 14 will discuss the microwave alternative). Fiber strands have a loss that, at the optical wavelengths used by cable television systems, is approximately two orders of magnitude lower than the best coaxial cables. Furthermore, the noise and distortion added by a typical 5- to 15-mile-long fiber-optic link is less than that generated in a comparable-length coaxial trunk line section.

This chapter will begin with a brief introduction to some principals of optics that are important to fiber transmission of light. Then optical fibers will be examined in detail, particularly the interaction between the fiber and the signals passing through it. Passive optical devices, comparable with their coaxial counterparts, will be treated next. Active devices will include transmitters, amplifiers, and receivers. Finally, the total performance of optical links will be examined with end-to-end performance calculations of noise and distortion. The unique characteristics of multiwavelength systems will be covered in Chapter 13.

12.2 Optical Basics

Light "waves" are just much higher-frequency RF waves — about 300,000 times the highest frequency used in coaxial cable systems and 20,000 times the microwave frequencies sometimes used to relay cable television signals. Optical wavelengths are very short (remember that $\lambda = c/f$ in a vacuum). For instance, visible yellow light has a frequency of about 5.5×10^{14} Hz, and thus a wavelength of only 0.00000055 m, more easily expressed as 550 nm (one nanometer is one billionth of a meter), considerably less than 1/10,000 of an inch.

At these frequencies, electromagnetic waves behave in ways that are sometimes counterintuitive. For one thing, physicists have long understood that the flow of energy is not continuous, but rather travels in "bundles" called *quanta*. In the case of light (and, for that matter, any radio waves), the carriers of the energy are massless subatomic particles called *photons* that travel, not surprisingly, at the speed of light (about 3×10^8 meters/second or, if you prefer, about 11.7 inches per nanosecond in a vacuum). The energy carried by each photon is directly proportional to the frequency of the signal, so photons at light frequencies are much more energetic than those at VHF radio frequencies. As a result, the "granularity" of VHF power transmission is extremely fine, whereas it is sufficiently coarse at optical frequencies to add measurable noise to optical transmission links.

So, is light energy carried by electromagnetic radio waves or by particles called photons? Fortunately, we do not have to answer that question. In fact, they are both convenient mathematical models that explain the behavior of light under various conditions. Physicists use whichever model is most convenient to the situation at hand, and so will we.

Also, fortunately, we need to know but little of the broad subject called optics to understand the use of light as a carrier of signals through glass fibers. We do, however, need to understand the fundamentals of the interaction between light and materials.

12.2.1 Velocity of Propagation and Wavelength

RF signals propagate more slowly through coaxial cables than in a vacuum. Specifically (Equation (10.8)), the ratio of cable propagation velocity to that in free space is $1/\sqrt{\varepsilon}$, where ε is the relative dielectric constant of the cable dielectric.

Given that light waves are simply higher-frequency radio waves, we would expect a similar phenomenon to occur in optical transmission lines, and it does. The difference is that the dielectric constant, at optical wavelengths, varies with frequency. At RF frequencies and below, the dielectric constant is dominated by molecular and atomic effects. At frequencies above the natural resonances of the molecular structures, they have less effect. The result is that the effective dielectric constant is both lower and wavelength dependent.

Rather than speak in terms of dielectric constant, the term *index of refraction* is used. The index of refraction of a material is numerically equal to the square root of the effective dielectric constant *at the optical frequency being considered*.

The speed of propagation through a material can be expressed as

$$v = \frac{c}{n} \tag{12.1}$$

where

v = the velocity of the propagated signal in meters/second
c = the speed of light in a vacuum (3×10^8 meters/second)
n = the material's index of refraction (≥ 1).

The slowing of the light waves through materials is the result of a phenomenon known as *forward scattering*. When light signals travel through material, it interacts with the individual atoms. The wave model of what happens is that individual electrons get "excited" into a higher energy state and then give up that energy by generating a wave of their own. The particle model is that once in a while a photon slams into an electron, exciting it to a higher energy state. After a while, the electron returns to its former energy level, emitting a photon of equal energy in the process.

Whatever the model, the result is the same: Energy from the transmitted wave is temporarily transferred to the material and then released in the form of new waves at the same frequencies. In the uniform-density, almost-pure glass used for optical fibers, most of the new waves are transmitted in the same direction as the original wave, but delayed by the time period between excitation and reemission (when waves are retransmitted in different directions or with varying delays, the total transmitted energy in the main wave is reduced, as will be discussed later). The new waves vectorially combine with the original wave to form a composite transmitted wave that is slightly delayed from the original. This new wave then interacts with material further along its path, adding further delay. Since this delay is proportional to the distance the wave is transmitted through the material, the result is a uniformly slower transmission speed. The index of refraction is how we express the resultant velocity relative to that in a vacuum.

The wavelength of the signal is related to the frequency, the velocity of propagation, and the index of refraction by

$$\lambda = \frac{v}{f} = \frac{\lambda_0}{n} = \frac{c}{fn} \tag{12.2}$$

where

$\lambda_0 =$ the wavelength in a vacuum
$f =$ the frequency in hertz

12.2.2 Reflection and Refraction

When light strikes a smooth, nonopaque surface, we know from common experience that some may be reflected and some may be transmitted into the material. We also observe that when light strikes a piece of glass or a still water surface at a very low angle, it is nearly all reflected, and the material resembles a mirror. At greater angles, only a small portion is reflected, and the material appears "transparent." These effects are due to the difference in index of refraction between glass and air.

Specifically, if we draw a general diagram (Figure 12.1) showing light propagating from a material on the left with diffraction index n_1 and hitting a perfectly smooth, flat interface with the material on the right (diffraction index n_2) at an angle from normal, we can express the equations governing the reflected and transmitted waves as follows:

- The transmitted wave will leave the interface at an angle ϕ relative to the normal such that

$$\frac{\sin\theta}{\sin\phi} = \frac{n_2}{n_1} \tag{12.3}$$

Thus, if $n_2 > n_1$ (as in the figure), the transmitted wave will be closer to normal relative to the interface plane, whereas if $n_2 < n_1$, it will be more parallel to the interface. This relationship is known as Snell's law.

In particular, if $\sin\theta > n_2/n_1$, there is no transmitted wave, and all the energy is reflected (this is strictly true only if the n_2 material is thick). This is a crucial condition for some types of fiber transmission, as will be seen. Since the value of the sine function cannot exceed 1, total reflection is possible only when $n_1 > n_2$ (the opposite situation from that shown in Figure 12.1). The angle where total reflection occurs (the "critical angle") $= \sin^{-1}(n_2/n_1)$.

- The reflected wave will leave the surface at an angle exactly opposite that of the incident wave, that is, an at angle $-\theta$ from normal (we will see later that this effect is used to advantage in certain types of optical fiber connectors).

The division of power between the reflected and refracted (transmitted) waves is a function of the refractive indices, the angle of incidence, and the direction of the electric fields of the incident signal. In the special case where the signal is normal to the interface, the ratio of reflected to incident power is

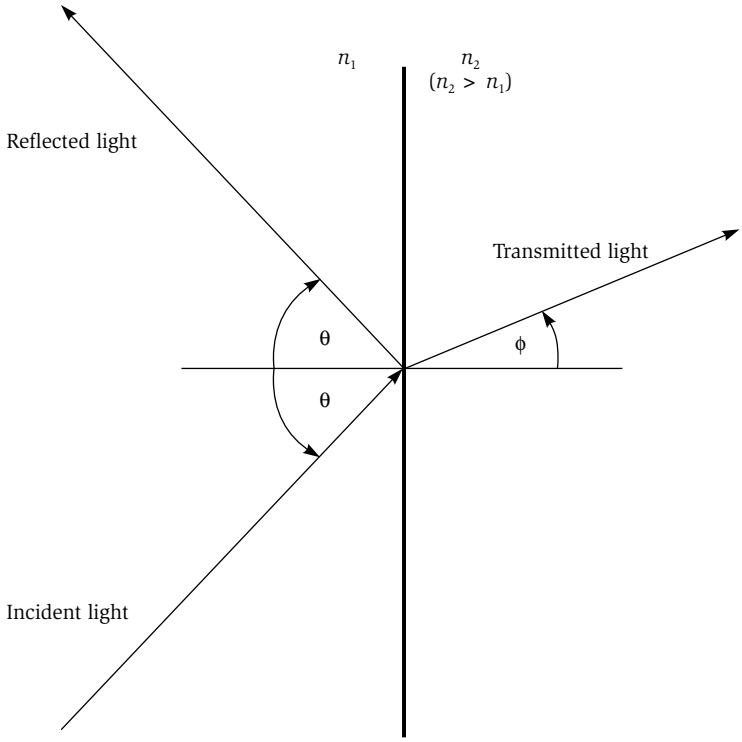


Figure 12.1 Light reflection and transmission at the interface between two materials.

$$\text{relative reflected power} = \left(\frac{n_2 - n_1}{n_2 + n_1}\right)^2 \tag{12.4}$$

and the ratio of transmitted to incident power is

$$\text{relative transmitted power} = \frac{4n_2n_1}{(n_2 + n_1)^2} \tag{12.5}$$

or in decibel terms more useful for transmission analysis:

$$\text{return loss (dB)} = 20\log\left(\frac{|n_2 - n_1|}{n_2 + n_1}\right) \tag{12.6}$$

$$\text{transmission loss (dB)} = 10\log\left[\frac{4n_2n_1}{(n_2 + n_1)^2}\right] \tag{12.7}$$

Thus, perfect power transmission and zero reflected power require that $n_1 = n_2$.

As an example of less-than-perfect transmission, air has an index of approximately 1 (actually about 1.003) and common glass an index of about 1.5 at visible light frequencies, so about 96% of light striking window glass at a right angle is transmitted and 4% is reflected, a transmission loss of about 0.18 dB. These relationships are important to the design of optical connectors.

12.2.3 Light Absorption

As with coaxial cables, some of the energy in a light signal propagating through materials is converted into heat. Though an exact understanding of the mechanisms is not essential, the effect is that an optical transmission path of uniform characteristics has a loss that, as with coaxial cables, is linear (in decibels) as a function of distance.

Absorption, if it is due to atomic resonance effects, is very wavelength dependent. At the wavelengths used for signal transmission through fibers, resonances effectively prevent the use of some of the optical spectra, whereas the effects of the resonances cause transmission characteristics at other wavelengths to be enhanced.

12.2.4 Scattering Loss and Rayleigh Scattering

Although on-axis (forward) scattering is what happens most of the time, photons may sometimes be emitted in different directions or with varying delays, leading to signal loss in the principal transmitted wave. Many mechanisms lead to this scattering, including impurities in the material, nonuniform densities, and nonlinearities due to extreme power levels. One such scattering phenomenon is known as *Rayleigh scattering*, after Lord Rayleigh, who is credited with first demonstrating that the sun's light is scattered by nitrogen and oxygen molecules in the air. Rayleigh scattering is very wavelength dependent, varying approximately as the inverse fourth power of wavelength, which is why the blue (short wavelength) end of the solar spectrum is scattered much more than the red end so that we see a sharp-edged red-yellow sun surrounded by a diffuse blue sky.

The loss of energy in light waves transmitted through standard optical fibers is due primarily to scattering loss and secondarily to absorption and several additional minor effects.

12.2.5 Wavelength Dispersion

Except for wavelengths close to the natural atomic resonance frequencies of the material, n varies with wavelength approximately as

$$n = A + \frac{B}{\lambda^2} + \frac{C}{\lambda^4} \tag{12.8}$$

where A, B, and C are constants related to a particular material. This wavelength dependence is known as *material dispersion*. Near atomic resonances, n varies wildly, and a significant amount of absorption also takes place. Such a resonance happens to occur near the wavelengths used for optical signal transmission.

A common example of dispersion in the visible spectrum is the common glass prism (see Figure 12.2). When white light strikes a piece of glass at an angle, signals at the various colors bend *toward* the normal (right angle to the surface) at different angles, due to the difference in n across the visible light spectrum (whose wavelengths range from about 400 to 650 nm). If the glass is a flat sheet, such as a window, the reverse bending takes place on the far side, and all signal wavelengths end up restored to their original angles, and we don't notice anything. If the sides are not parallel, however, then when the light impinges upon the second interface, the angular deviation is made even greater as each wavelength bends *away* from the normal by varying amounts depending on the difference in n as a function of wavelength, and we observe the separated spectrum.

Dispersion is a major concern in optical communication cables, not primarily because of the bending effect, but because of the differing transmission speeds associated with changing n. If the components of a signal are, for some reason, generated over a range of wavelengths and transmitted through a cable whose index of refraction varies, then the components will not arrive simultaneously at the receiver. Depending on the signal format, this can lead to various kinds of distortion. The affect of dispersion is, in fact, just optical group delay (see Chapter 15) and has the same types of effect.

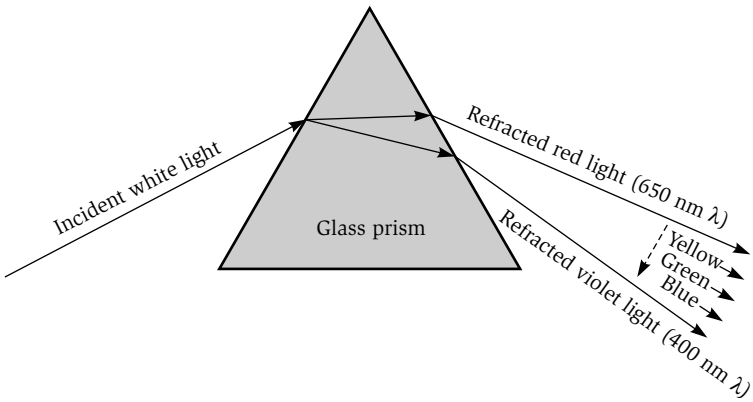


Figure 12.2 Chromatic dispersion of visible light in a glass prism.

In addition to the foregoing, the interaction of light with very small glass fibers produces some unique interactions that will be discussed in the next section.

12.3 Multimode Optical Fibers

If we construct a long, smooth, narrow (but large compared with the wavelength of the light), circular-cross-section bar of glass and shine light into one end, some of the light will travel along the length of the bar and exit the far end. That will be true even if the bar isn't perfectly straight. The reason can be seen in Figure 12.3.

Since the refractive index of glass is greater than that of air, light entering the bar at a sufficiently shallow angle will be completely reflected at the glass-to-air interface and will stay in the bar, simply bouncing along from side to side and eventually exiting the far end. Light entering at a greater angle will lose some of its energy at each reflection and will quickly be dissipated. The maximum angle where all of the energy is internally reflected defines the circular *cone of acceptance* for the bar.

If the bar is curved slowly (again, relative to the wavelength of the light), the light will still be totally reflected and will follow the curvature. If it is curved too tightly, however, some of the light will strike the outer surface at angles above the critical angle, and a percentage of the energy will be diffracted out of the bar.

The performance of this simple transmission line can be improved (though the cone of acceptance will be much smaller) if we coat the outside with a second layer of glass that has a slightly lower index of refraction than the core. So long as this coating is sufficiently thick, we can now handle the rod or place it next to other objects without interfering with its light transmission properties.

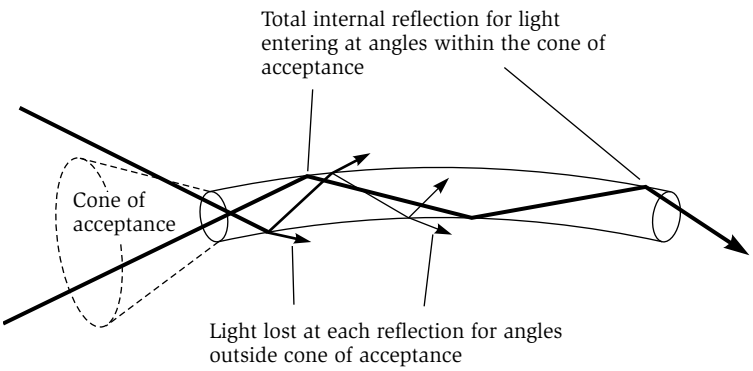

Total internal reflection for light entering at angles within the cone of acceptance
Cone of acceptance
Light lost at each reflection for angles outside cone of acceptance

Figure 12.3 Light transmission through a bent glass bar.

If the entering light is imperfectly collimated (that is, it enters at more than one angle), it can readily be appreciated that the light traveling straight down the core will go less distance than light that is near the critical angle and bounces frequently. The difference in transmission path length will depend on the length of the bar and causes another form of dispersion (known as *modal dispersion*) that is extremely important to signal transmission.

If we modify the index of refraction so that it decreases radially from the center, light that travels at a small angle with respect to the axis will slowly be bent back toward the core as it travels farther from the center. It will, however, have traveled more quickly than light transmitted straight down the core, because it will have spent a percentage of its travel time in a medium with a lesser n and therefore greater velocity. Thus, by varying the index of refraction properly, the variation in transmission time with respect to angle can be reduced.

Now imagine that the "bar" of glass is round, with a core whose index of refraction is 1.49, whose diameter is only 50×10^{-6} or 62.5×10^{-6} meters (50 or 62.5 μm, or "microns") and whose length is up to thousands of feet, and you have a typical "multimode" optical fiber. The preceding analysis, with a few exceptions, explains how signals are transmitted in such cable. Fibers made with just a single coating are known as *step index* fibers, whereas those with a graduated index of refraction are known as *graded index* fibers.

Where the simplified analysis breaks down is that when the core of a fiber is not very large compared with the optical wavelength, only certain transmission paths are supported, whereas under the simplified analysis, light introduced at any angle within the cone of acceptance would be transmitted. The most common wavelengths transmitted through multimode fiber are 850 and 1310 nm. In glass with $n = 1.49$, these wavelengths shorten to 537 and 879 nm, respectively. Thus the diameter of a 62.5-μm fiber core is 116 or 71 wavelengths: large but not infinite.

Little or no multimode fiber is used for cable television long-distance signal transmission because of the angle-dependent dispersion effect discussed earlier. This form of dispersion is caused by the difference in axial transmission speeds of light traveling along different transmission paths or modes and is known as *modal dispersion*. How dispersion affects signals is illustrated by Figure 12.4. If a short burst of imperfectly collimated monochomatic (all one frequency) light is introduced into multimode fiber at time T0, the signal will be introduced into the fiber at a range of angles (relative to the fiber axis) and, thus, travel different total distances. If the pulse is examined after it has passed through a length of fiber, we will "see" first a little light from the signal traveling parallel to the core, then from the next fastest modes, and so on, until we finally receive light from the highest angles. Similarly, at the end of the pulse, the light from the fastest modes will disappear first, and then gradually all the light will disappear. The effect is to shorten the top of the pulse and broaden the bottom (T1). If the pulse is sufficiently short, or the fiber sufficiently long, the top will disappear entirely (T2). For still longer fibers, the leading and trailing edges of the pulse will

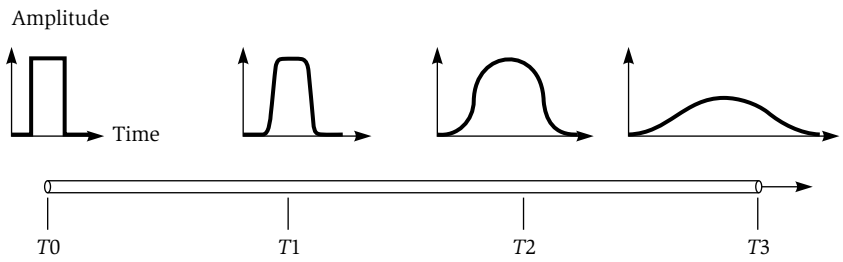


Figure 12.4 Pulse degradation through multimode fiber.

overlap, and we will see a longer, lower-amplitude pulse (T3). If we transmit a series of pulses (representing, for instance, a digitally encoded signal), then the maximum pulse rate will be determined by when the residual effects from one pulse interfere with the onset of the next pulse. One way of expressing this maximum is the bandwidth of the fiber. If we drive the fiber with light that has been amplitude modulated at varying rates, the bandwidth is the point where the amplitude of the demodulated signal drops by 3 dB relative to its value at low modulation frequencies. Alternatively, the bandwidth is approximately equal to $0.35/t_r$, where t_r is the apparent rise time of the detected pulse when the input pulse has zero rise time.

Multimode fibers are rated according to the amount of this length-dependent bandwidth. Typical fibers with a 62.5-micron core diameter exhibit a bandwidth–length product of 300 MHz-km at 1310 nm and 160 MHz-km at 850 nm. Fibers with a smaller, 50-micron core can reach 400 MHz-km. What this means is that a 50-micron core fiber that is 1 km long will have a bandwidth of 400 MHz, whereas one that is 10 km long will have a bandwidth of only 40 MHz.

Clearly, such fibers are not suitable for transporting 870 MHz of cable television spectrum over the long distances required in modern networks.

12.4 Single Mode Optical Fibers

If the core diameter is reduced sufficiently, fibers will support only light traveling collinearly with the axis (known as the LP_{01} mode), thereby eliminating modal dispersion. Such fibers, known as *single mode*, are almost universally used for long-distance, high-bandwidth applications.

12.4.1 Structure

The most common single mode fiber construction consists of an 8.3-micron-diameter core, surrounded by *cladding* glass with a uniform, lower index of

refraction and extending out to about 125 microns. This, in turn, is surrounded by a protective layer (the *coating*) extending out to 250 microns. This outer layer is generally colored to enable easy fiber identification in bundles. This construction is known as *matched-clad*, *nondispersion-shifted* fiber.

The necessary condition for suppression of higher-order modes is

$$V = \frac{2\pi a\sqrt{n_1^2 - n_2^2}}{\lambda} < 2.405 \tag{12.9}$$

where

V = the V parameter, or normalized frequency
a = the core diameter in micrometers
n_1 = the core refractive index
n_2 = the cladding refractive index
λ = the free-space wavelength in micrometers[1]

As can be seen, this requires very small core diameters and a very small difference in the refraction coefficients for the core and the cladding.

A variant on the preceding construction is to use two layers of cladding: a lower-n inner cladding followed by an outer layer with an intermediate n. This is known as *depressed clad*, *nondispersion-shifted* fiber and is otherwise the same in construction as matched clad fiber. With few exceptions, its performance as a transmission medium is very similar as well. Figures 12.5(a) and (b) show the refraction index profile and layers of construction of both types.

With either type, the light travels mostly in the core but also partially in the cladding. The effective index of refraction is about 1.467 for either type at either of the common wavelengths used: 1310 nm or 1550 nm.

12.4.2 Attenuation

The total loss due to absorption and scattering is amazingly low. At 1310 nm, typical guaranteed losses for cables sold to the CATV market are only 0.35 dB/km, whereas at 1550 nm, they drop to 0.25 dB/km. Otherwise stated, at 1550 nm, only half the light energy is lost through a strand of glass 12 km (about 8 miles) long.

The curve of attenuation-versus-free-space wavelength is shown in Figure 12.6 (also from Ron Cotton). The local minima around 1310 and 1550 nm are known as the second and third windows of transmission (the first is at 850 nm). The high absorption peak around 1380 nm, known as the *water peak*, is due to absorption by hydroxyl ions. The water peak height has decreased as fiber production techniques have improved, and at least one manufacturer now offers fiber entirely free from this effect, as shown in Figure 12.7.[2]

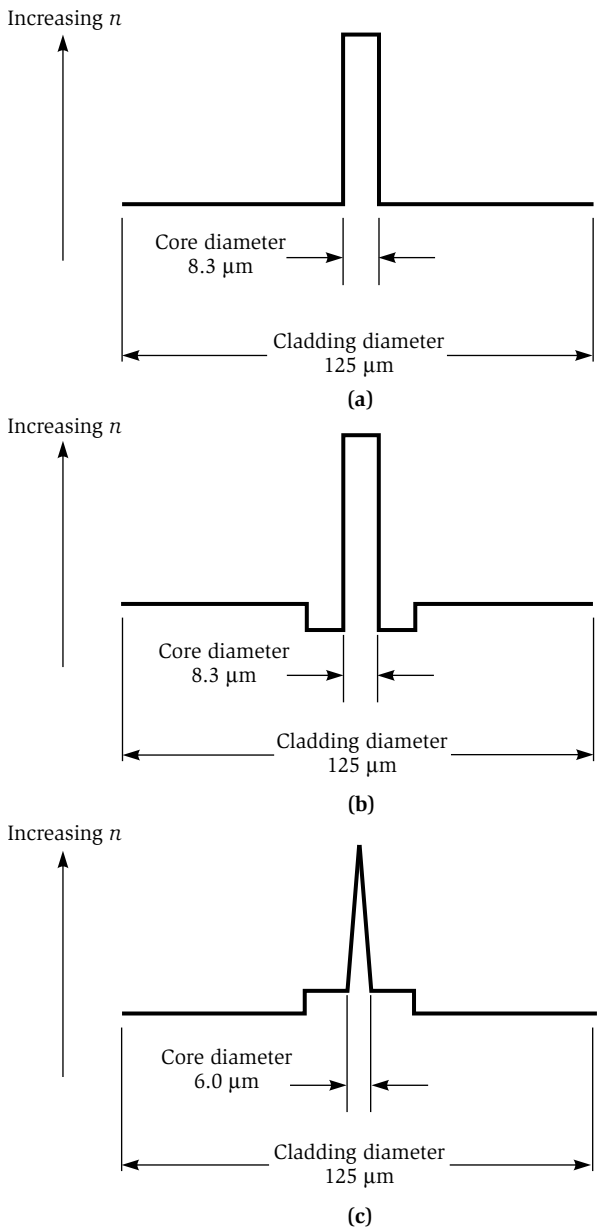


Figure 12.5 Cladding profiles for various fiber types. (a) Nondispersion-shifted matched clad fiber doping profile. (b) Nondispersion-shifted depressed clad fiber doping profile. (c) Dispersion-shifted fiber — one of several possible doping profiles.

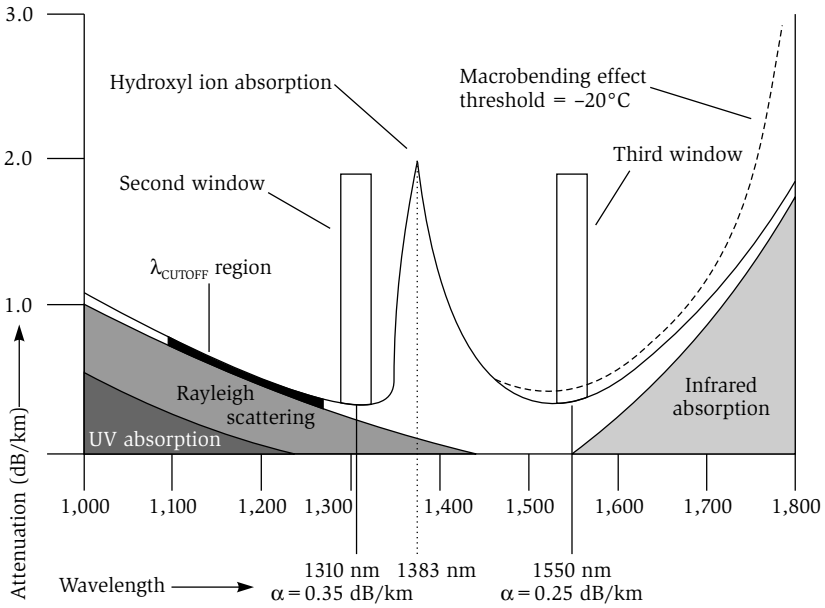


Figure 12.6 Fiber-optic loss as a function of wavelength.

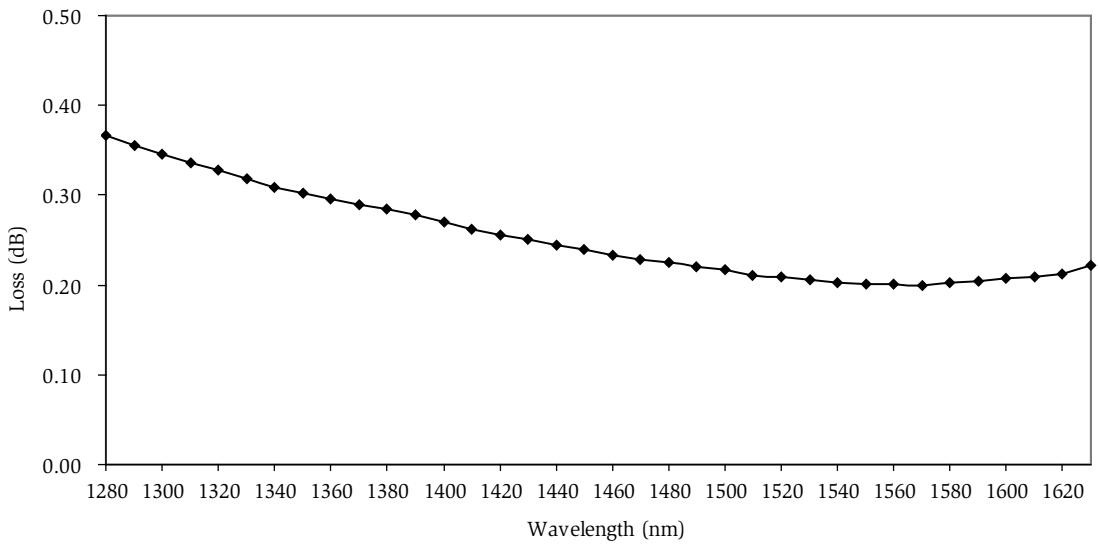


Figure 12.7 Typical ZWP attenuation curve.

12.4.3 Chromatic (Wavelength) Dispersion

With the large modal dispersion eliminated, more subtle dispersion mechanisms become the limiting factors. *Chromatic dispersion* is a measure of the degree to which the effective propagation velocity changes as a function of wavelength. It is the sum of two factors: *material dispersion*, which is a measure of the change in refraction index of the glass with wavelength, and *waveguide dispersion*. In optical fibers, the signal travels partially in the core and partially in the cladding, and the total *mode field diameter* changes with wavelength. Since the refraction index is different in the core than in the cladding, a change in mode field diameter also results in a change in average dispersion index and, therefore, signal velocity. The ratio of wavelength change to velocity change due to this effect is known as *waveguide dispersion*.

As with modal dispersion, chromatic dispersion is a linear function of transmission system length. The units of chromatic dispersion are picoseconds per nanometer-kilometer; that is, for a 1-nm free-space wavelength change, this gives the number of picoseconds of delay change per kilometer of fiber length.

Standard fiber exhibits zero chromatic dispersion near 1310 nm (because the slopes of waveguide and material dispersion components are equal in magnitude and opposite in direction at that wavelength). In cases where it is important to have low dispersion at 1550 nm, the null point can be shifted upward by altering the fiber doping profile and/or using several layers of cladding. Figure 12.5(c) gives one example of a doping profile for dispersion-shifted fiber.[3] Figure 12.8

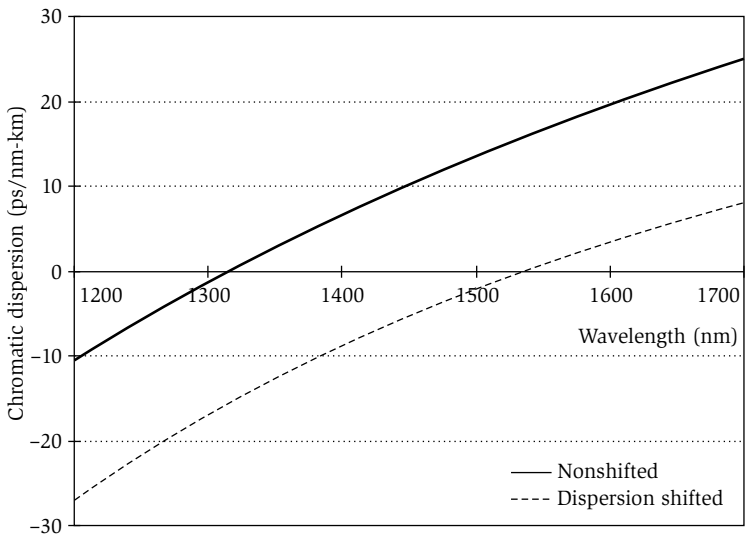


Figure 12.8 Chromatic dispersion for standard and dispersion-shifted fiber.

shows a typical chromatic dispersion curve for both non-dispersion-shifted and dispersion-shifted single mode fibers.

Typical commercial specifications for chromatic dispersion are 2.8–3.2 ps/nm-km from 1285- to 1330-nm wavelength and 17–18 ps/nm-km at 1550 nm for nonshifted fibers.[4]

Dispersion is very important in communications circuits because the optical sources used do not transmit on a single wavelength. The interaction between various source wavelengths and network performance will be discussed later in this chapter.

12.4.4 Polarization Mode Dispersion

Light transmitted through single mode fiber may be thought of as two separate signals (polarization modes) with their electric fields 90° apart relative to the axis of the fiber. So long as the fiber looks exactly the same to both signals, they will have the same transmission time and will arrive in phase at the detector.

Two conditions, however, can detract from equal transmission velocity: dimensional irregularities and nonequal index of refraction.

- As discussed, waveguide dispersion is a function of the mode field diameter, because the signal travels partially in the core and partially in the cladding. Should the fiber core not be completely round, the ratio of core to cladding utilized for one polarity will be different from that for the other.
- More subtly, the glass may have a dimensional preference in its atomic structure that causes n to be greater in one direction than the other. This preference is typically nonuniform along the fiber and also modified by local stresses and temperature variations.

The situation is made more complicated because energy is exchanged between the polarization modes due to imperfections in the fiber, connectors, splices, and so on.

The net effect of all these mechanisms is called *polarization mode dispersion* (PMD). PMD is typically wavelength dependent and interacts with optical transmitter wavelength changes to cause system distortion. A typical commercial specification for PMD is $\leq 0.5\,\text{ps}/\sqrt{\text{km}}$ at 1310 nm in non-dispersion-shifted cable,[5] though typical values are $0.1\text{–}0.2\,\text{ps}/\sqrt{\text{km}}$. At these values, PMD does not add materially to total system distortion in most cable networks.

12.4.5 Attenuation in Bent Fibers

As with the waveguides utilized for microwave signal transmission lines, single mode fibers may not be sharply bent without causing signal loss. Two classifications of bends are normally considered. *Microbends* are microscopic imperfections,

often introduced in the manufacturing process, such as irregularities in the core material or in the interface between core and cladding. Losses due to microbending are included in manufacturers' specifications for cable performance.

Macrobends are bending of the completed fibers. If the radius of a macrobend is sufficiently small, significant amounts of the light will be transmitted out through the cladding and be lost. For that reason, manufacturers specify cable and individual fiber-bending radii that must be respected.

The fact that light is not constrained to remain within the fiber when bent has been exploited by manufacturers of light insertion and detection (LID) fusion splicers. They bend the incoming fiber sufficiently to insert light on one side of the splice location (light can "leak" into, as well as out of, fiber when it is bent) and then bend it on the other side and detect the leaked light. By manipulating the position of the fibers at the joint for maximum transmission before fusing, splice loss can be minimized. One ambitious manufacturer even reached the field trial stage with a fiber-to-the-curb (FTTC) telephone and cable television distribution system that used bent-fiber couplers to extract and insert light at every node.[6] Their first version utilized multimode fiber, but later versions utilized single mode fiber.

A typical commercial fiber-bending specification is 0.05–dB maximum loss @ 1310 nm and 0.10 dB @ 1550 nm for 100 turns of fiber wound around a 75-mm- (2.95-inch-) diameter mandrel. The specified loss for a single turn around a 32-mm- (1.26-inch-) diameter mandrel is 0.5 dB @ 1550 nm.[7] In general, bending losses are higher at 1550 nm than at 1310 nm.

12.4.6 Stimulated Brillouin Scattering

When the power level of monochomatic light transmitted into a long fiber strand is increased, the output power increases proportionately until a threshold is hit. Beyond that level, the received power stays relatively constant, and energy reflected toward the source increases dramatically. The signal-to-noise and signal-to-distortion ratios of the received signals both degrade. This phenomenon is known as *stimulated Brillouin scattering* (SBS).

SBS occurs because glass is electrostrictive — that is, an electrical field causes a mechanical stress on the material. When the field associated with the light signal reaches a certain amplitude, only some of the energy from excited electrons is returned as forward-transmitted waves. The remainder is translated into an acoustic wave that propagates through the material. This acoustic wave, in turn, modulates the index of refraction due to the sensitivity of the index of refraction in glass to pressure. The varying index of refraction then causes the main light wave to alternately slow down and speed up, causing the detected signal to have increased distortion. A portion of the reflected light is re-reflected at acoustic wavefronts due to the change in index of refraction, and the random phase relationship between the original and double-reflected signal, as received at the detector, causes increased noise.

SBS is a function of fiber geometry, loss, and length, as well as transmitted power. It is also a wavelength-specific process, in that energy at light frequencies separated by as little as 20–100 MHz behave independently, so when light is transmitted at several wavelengths each experiences about the same Brillouin threshold.

The basic equation for the single-wavelength SBS threshold is

$$P_{SBS} = \frac{21A\alpha}{\gamma(1 - e^{-\alpha L})} \tag{12.10}$$

where

P_{SBS} = the threshold power of SBS scattering in mW

A = the cross-sectional area of the fiber core (approximately 80×10^{-12} m^2 for non-dispersion-shifted fibers and 50×10^{-12} m^2 for dispersion-shifted fibers)

γ = the Brillouin gain coefficient ($2-2.3 \times 10^{-14}$ meters/mW, depending on the degree of polarization randomness in the signal)

α = the fiber's attenuation in dB/km $\times$ 2.3×10^{-4} (numerically equal to twice the attenuation in nepers/m)

L = the fiber's length in meters[8]

Figure 12.9 shows typical SBS threshold values for standard (Std) fibers at both 1310 and 1550 nm and for dispersion-shifted (DS) fibers at 1550 nm.

Power higher than the monochromatic SBS threshold can be transmitted if divided among optical carriers whose frequencies differ by more than the SBS linewidth. Some vendors use spectrum-spreading techniques to allow transmitted powers as high as +17 dBm (50 mW) at 1550 nm in standard fiber without significant distortion. Clearly SBS extension becomes a balance between widening the spectrum sufficiently to avoid the Brillouin limit while keeping it sufficiently narrow to avoid significant chromatic dispersion. One method commonly used is to phase modulate the source with a sufficiently high amplitude and frequency to adequately spread the spectral power while creating added modulation sidebands that fall above the highest RF signal carried by the system.[9]

12.4.7 Self-Phase Modulation Interacting with Dispersion

Another power-related effect that occurs in optical fibers is known as *self-phase modulation* (SPM). This is caused by the fact that the index of refraction, n, is slightly modulated by variations in the instantaneous intensity of the transmitted light. This causes a different propagation velocity (as the instantaneous power varies) and thus an effective envelope distortion, resulting in phase modulation as seen at the detector. The nature of the distortion will be affected

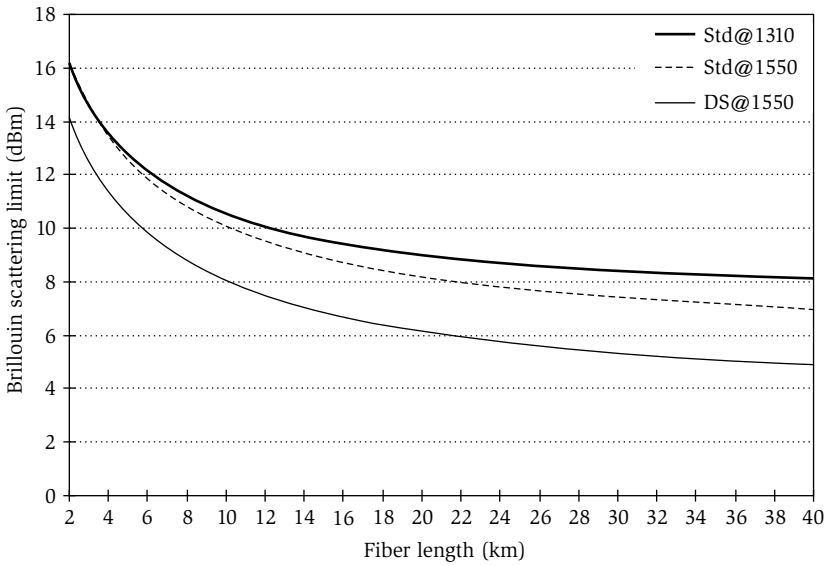


Figure 12.9 Brillouin scattering limit for monochromatic light as a function of fiber length, type, and operating wavelength.

by whether the wavelength is centered on the fiber's dispersion null or on one side of the null.

In general, the effect of SPM will be an increase in the level of composite second-order (CSO) distortion through the link. Although the analysis is not straightforward, the C/CSO will be degraded when

- The chromatic dispersion is higher.
- The fiber is longer.
- The RF carrier frequency is higher.
- The wavelength of the light is longer.
- The average light level is higher.
- The optical modulation index (OMI, the intensity variation due to the modulating waveform) per carrier is higher.
- The core diameter of the fiber is smaller.
- The fiber attenuation is lower.

Figure 12.10 shows the measured and calculated C/CSO degradation due to SPM as a function of average optical power level under the following conditions:[10]

- Length = 50 km of non-dispersion-shifted fiber with a power coefficient of diffraction. index equal to 2×10^{-20} m^2/W, attenuation of 0.0576 nepers/km, and chromatic dispersion at 1553 nm of 17 ps/nm-km.
- Channel loading = 77 RF carriers, with OMI/carrier approximately 2.5% (top curve) and 3.4% (lower curve).
- C/CSO measured on top channel (547.25 MHz).
- The transmitter is an externally modulated DFB laser at 1553 nm (optical transmitter characteristics are discussed later in this chapter).

In Figure 12.10, the solid lines are the calculated results, and the discrete points are measured data. The referenced paper offers no explanation for the deviation between calculated and measured data at higher OMI values.

Figure 12.11 shows the variation as a function of length for a fixed +18-dBm optical power level. For lengths greater than 50 km, the signal was amplified after the first 50 km and transmitted through the additional fiber length. The solid line is the calculated performance, while the × marks are measured data.

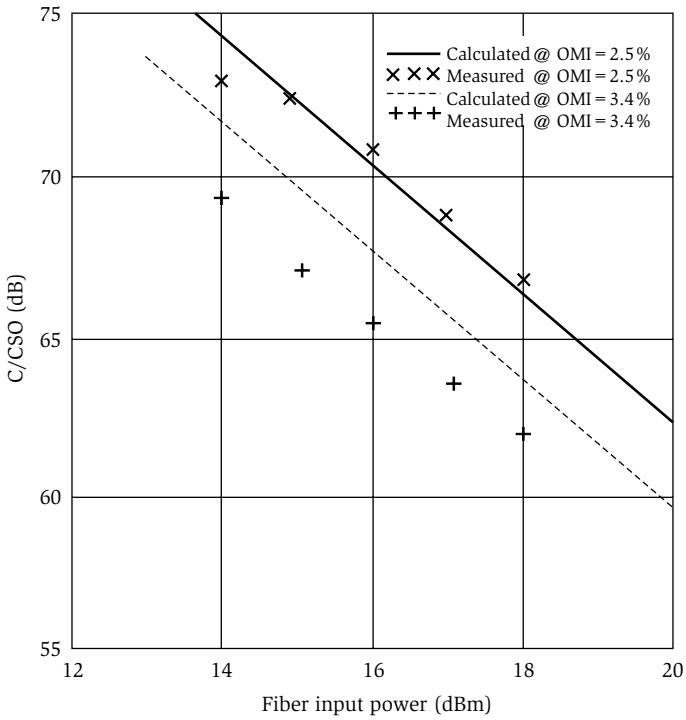


Figure 12.10 CSO degradation due to self-phase modulation.

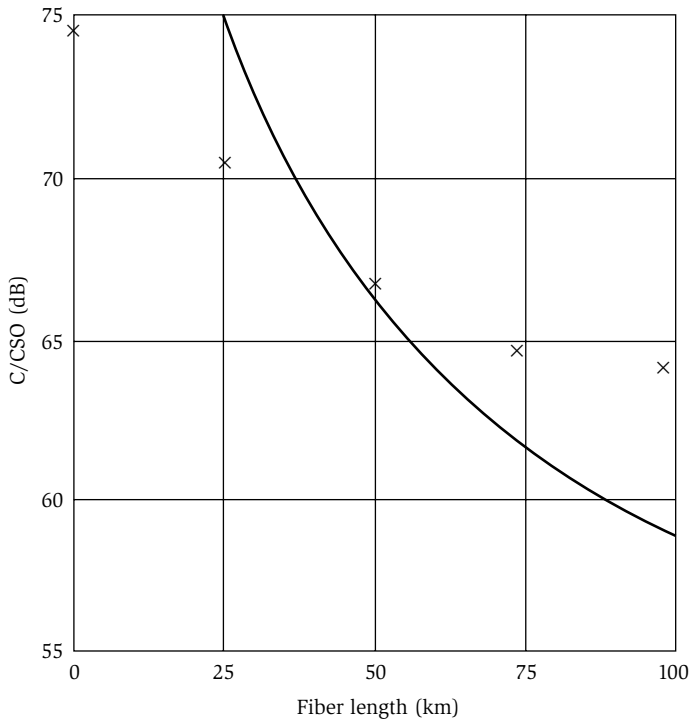


Figure 12.11 CSO degradation due to SPM as a function of fiber length.

As can be seen in these graphs, at higher optical levels, the degradation in C/CSO is significant. In some cases, this may be a compelling argument for the use of dispersion-shifted fiber for long 1550-nm analog links.

In systems containing components other than fiber, SPM also interacts with any response variation to produce additional distortion components. This is covered in more detail in Chapter 13, since wavelength division multiplexers are often the primary contributors to this additional CSO source. Because self-phase modulation effects the transmission properties of a link, it is one of the mechanisms that limits performance when more than one optical signal (on different wavelengths) shares a fiber, a phenomenon known as *cross-phase modulation*, which is discussed in Chapter 13.

12.5 Network Passives

In coaxial distribution systems, passive devices are used for various signal-splitting and -combining functions. Optical equivalents exist for many of the same

functions, with the obvious exception of multiplexing of 60-Hz power with signals. Discussed next are the essential passive components used in linear optical links.

12.5.1 Connectors

Methods for joining fibers fall into three general classes: fusion splices, mechanical splices, and connectors. *Fusion splicing* consists of carefully preparing the end of each of the fibers to be joined and then bringing the ends together carefully and literally melting them. Automated and semiautomated fusion splicers have evolved that consistently produce splices with less than 0.1 dB of loss.

Mechanical splices generally involve similar fiber-end preparation. The difference is that the fibers are brought into proximity and embedded in a gel-like material whose index of refraction closely matches that of glass. A mechanical sleeve holds the fibers in position. Though the loss at such splices is higher than at fusion splices (typically 0.1–0.2 dB), they are quick to manufacture and are favored for rapid field repair tasks in cases where a fusion splice crew can later replace the temporary fix.

A wide variety of connector families have been developed by various vendors. These are used for interfacing between active equipment and fiber lines and for patch panels that allow test points and reconfiguring of systems. Typical connector losses are higher than mechanical splices (about 0.25–0.5 dB) and are variable with remating.

The difficulty in improving connectors lies in their imperfect fiber-to-fiber interface. As discussed at the beginning of this chapter, signals lose approximately 0.18 dB at each fiber–air interface. Unless the fiber ends are in intimate contact, a mated connector pair will have two such interfaces and a loss of 0.36 dB due to the two fiber–air interfaces alone. Worse, the return loss from each interface could be as bad as 14 dB.

A number of different connector families have been used over the years, and more will likely be developed in the future. The SC series is currently popular in North America. This is a small push-on–pull-off connector that is capable of low insertion loss and good return loss if the fiber ends are properly polished. A zirconia ceramic ferrule is used to align the two fibers. It has a central capillary hole into which the fiber is cemented, following which the end is polished. Connectors may be attached in the field or supplied preattached to fiber pigtails that can be fusion spliced in the field.

Ultra-Polished and Angle-Polished Connectors

SC connectors, as well as some other series, either come with flat, highly polished tips or have angle-polished tips. The flat (SC/UPC) version allows for somewhat better insertion loss when cleaned properly but has typically 5-dB worse return loss than the angled (SC/APC) version.

The reason for the lower reflections from the APC version is shown in Figure 12.12. Whatever light is reflected at the fiber end (due to alignment errors

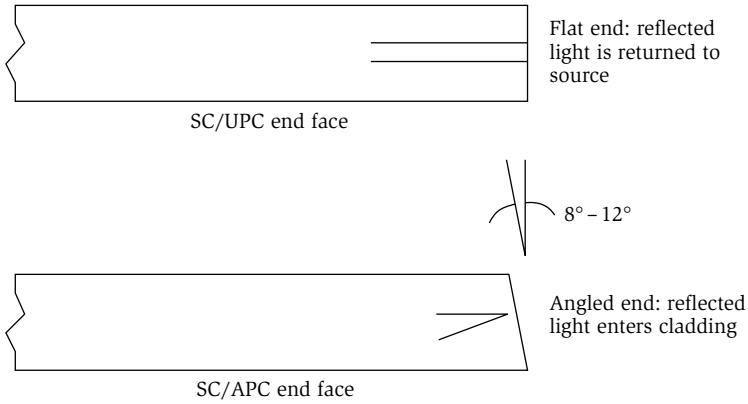

Flat end: reflected light is returned to source
SC/UPC end face
8° – 12°
Angled end: reflected light enters cladding
SC/APC end face

Figure 12.12 UPC versus APC connector.

between the fibers, air gaps, imperfect polishing, or differences in the index of refraction) is returned at an angle equal to twice the end angle of the fiber, which is 8° to 12°. The reflected light is outside the cone of acceptance of the fiber and so is not transmitted toward the source.

By contrast, the UPC connector requires a higher degree of polishing, which results in lower insertion loss, but with the disadvantage that whatever light is reflected will propagate back toward the source. Because of this difference, the return loss of an imperfectly cleaned UPC connector degrades more quickly than that of an APC connector.

The preoccupation with return loss levels in optical circuits, which are much lower than the 16 dB typical for coaxial components (as discussed in Chapter 10), is because directly modulated laser diode transmitters are very sensitive to reflections, as discussed in Section 12.6. Generally, optical return loss values must be significantly better than 40 dB to avoid problems.

Connector Cleaning Issues

Proper cleaning of optical connectors is mandatory for good performance. Many problems in optical links are caused by dirty connectors. Proper installation and maintenance practices call for careful cleaning and inspection of all connectors before mating or remating, preferably using inspection microscopes made for that purpose. Recent experiments have shown that, even when connectors on new equipment were carefully cleaned and capped for protection, they were unacceptably dirty when uncovered for use after undergoing a simulated shipping environment.[11]

If dirty connectors are mated, it is possible that the foreign material, if abrasive, can damage the polish of the fiber-end faces, permanently degrading the performance. In that case, the only remedy is to install a new connector.

Two methods of cleaning are currently accepted. In the first, a slightly abrasive tape is wiped across the connector end face. The tape is contained within a cartridge and is advanced after a set number of cleanings. This method has been found effective with SC/UPC connectors; however, its use on SC/APC connectors is questionable since excessive use can damage the angle polish.

The second method uses at least 91% pure isopropyl alcohol applied with a lint-free swab, followed by a dry swab and then blowing it dry using canned air. This method seems to be adequate for any type of connector. Standard "rubbing alcohol" is not acceptable because of impurity levels that may leave a residue.

12.5.2 Signal Splitters

Signal splitters can be made with virtually any splitting ratio. Excess losses (that is, the difference between input light power and the sum of the powers appearing at all output ports) are generally under 1 dB. Often, multiple splitters are packaged to create multiple port splitters with 10 or more legs.

12.5.3 Wavelength Division Multiplexers

WDM couplers are the optical equivalent of RF diplex filters. They are used to route light to the correct port on the basis of the optical wavelength. The simplest versions divide 1310-nm light from that at 1550 nm. More complex WDM couplers can separate multiple wavelengths in the 1550-nm range that differ in wavelength by as little as 0.8 nm. Such systems are known as *dense WDM*. Wavelength division multiplexers are discussed in greater depth in Chapter 13.

12.5.4 Attenuators

Optical attenuators are used both as test equipment and for absorbing excess light in some transmission links. Both fixed and variable units are available. A variety of techniques, including controlled bending of an internal fiber, are used to create the desired signal loss.

12.6 Linear Optical Transmitters

Transmitters generate optical carriers and intensity modulate those carriers with wideband RF spectra. In addition to linearity and noise requirements that are analogous to those placed on RF amplifiers, optical transmitters must exhibit optical characteristics that are compatible with the interconnecting fiber strands.

12.6.1 General Characteristics

In general, optical transmitters use an electrical signal to modulate the power of a light source. The most common application is simple on–off modulation used to convert high-speed binary signals to light pulses. At the receiving end, detectors need only detect the presence or absence of light to accurately reproduce the original digital stream. Digital modulation is a conceptually simple and robust format, and relatively high noise levels can be tolerated. At the present state of the art, speeds in excess of 10 gigabits per second (Gb/s) are commercially deployed, whereas optical losses of 30 dB can be tolerated between transmitters and receivers.

Figure 12.13 shows theoretical bit error rate as a function of signal-to-noise ratio for a baseband, binary transmission system.[12] Quite moderate S/N ratios (compared with those required for modulated analog NTSC signals) yield very good bit error rates. Baseband digital modulation is frequently employed in fiber-deep architecture (fiber-to-the-curb or fiber-to-the-home) due to the lower cost of the transmitters relative to those commonly used in HFC systems. Chapter 19 discusses fiber-deep architecture in detail.

By contrast, transmission of an FDM analog signal is much more complex. In the usual cable television application, the composite headend output waveform (with spectral components extending from 50 to 750 MHz or more) is used to control the light intensity in a proportional manner. The modulation must be highly linear or, more correctly, must complement exactly the demodulation characteristic

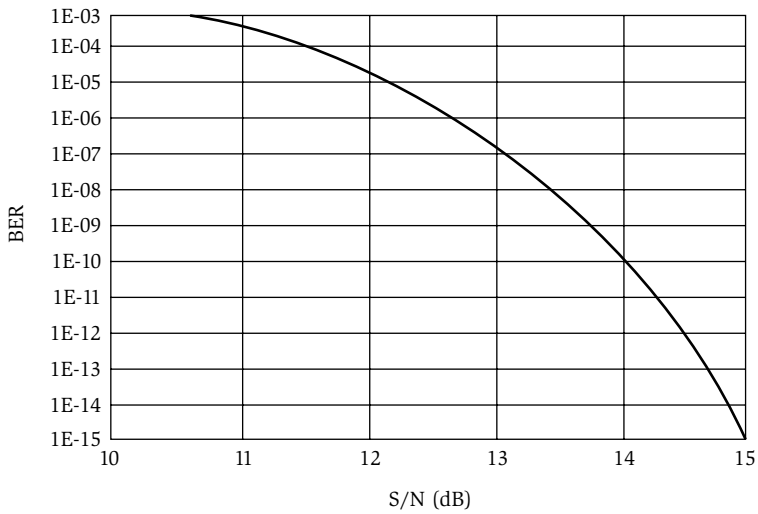


Figure 12.13 Bit error as a function of signal-to-noise ratio for a binary digital transmission system.

so that the total link is highly linear. Not only that, but the end-to-end C/N of each multiplexed RF signal must be compatible with signal-quality requirements. Typically, the fiber-optic link is required to provide a C/N of 50 dB or greater for each analog video carrier with composite distribution products 60–65 dB down. Since stimulated Brillouin scattering and other nonlinear fiber effects limit the maximum transmitted power to +10 to +17 dBm, and receiver power levels of about 0 dBm are required to sufficiently overcome "shot noise" (see Section 12.8) at the receiver, this limits practical optical budgets for typical 80-channel fiber-optic links. These relationships will be further explored later in this chapter.

In fact, optical transmitters and receivers are not linear devices but "square law" devices; that is, the instantaneous light output *power* of a transmitter is proportional to the input *current* and thus to the square root of the input signal power. At the other end of the circuit, the RF output power from the detector is proportional to the square of the optical power received, so the total link is nominally linear (predistortion is often used to overcome residual nonlinearities). As will be seen, however, the square law transfer function has an effect on noise and distortion addition. In particular, because of the square law detector transfer function, a change of 1 dB in optical loss will result in a 2-dB change in detected RF power, leading to the commonly stated, but incorrect, statement that "optical decibels are twice as big as RF decibels."

From a system standpoint, and regardless of the technology used to generate the signals, transmitters can be characterized by their wavelength, spectral purity, and stability and by their bandwidth, linearity, and noise levels. These interact with the fiber and detector characteristics to determine the total link quality.

12.6.2 Directly Modulated Fabry–Perot Laser Diodes

Although a number of technologies can produce the roughly monochromatic light required for optical transmitters, few meet the dual requirements of sufficient power output along with a linear relationship between input voltage or current and light output power.

One that does is the Fabry–Perot semiconductor laser, named for the French scientists Charles Fabry and Alfred Perot. In a Fabry–Perot (F–P) laser, light is reflected and re-reflected between two "mirrors" at either end of a semiconductor material that has been biased electrically. The material and two mirrors form a resonant cavity that roughly determines the wavelength of the light produced. One of the mirrors is only partly reflective, allowing some portion of the light to "leak" out into an external fiber, whereas most is internally reflected. This is directly analogous to a "high-Q" resonant L–C tuned circuit where the circulating energy is much higher than that coupled into a load.

Physical processing limitations dictate that the spacings between the mirrors form a resonant chamber long enough that oscillation can take place at any of several frequencies (longitudinal modes) for which the cavity length is an integer number of half wavelengths. Though the total output power of F–P

diodes is relatively constant (for a given bias current), that power fluctuates randomly among several modes, each with a slightly different wavelength. For typical 1310-nm F–P diodes, the spectral lines are spaced about 1 nm, and the energy is spread among several such lines. When this signal is transmitted through a fiber, light at each of these frequencies transmits at a slightly different speed due to chromatic dispersion. At the detector, the randomly varying transmission times translate into noise.

In addition to this *mode partition noise*, F–P diodes exhibit a certain amount of instability in their total optical power, known as *relative intensity noise* (RIN), and typically deviate slightly from an ideal square law transfer curve. For all these reasons, F–P diodes are not generally used for wideband downstream transmission, although their moderate cost makes them attractive for upstream use, where, typically, a single analog video channel or a limited number of digitally modulated signals are carried.

12.6.3 Directly Modulated Distributed Feedback (DFB) Laser Diodes

A modification of the basic F–P laser uses a diffraction grating (which serves as an optical tuned circuit) along the length of the cavity to restrict oscillation to a single mode. Known as a *distributed feedback* (DFB) laser, this is the most common light source used in downstream optical transmitters. High-grade DFB lasers may have linewidths (the width of the optical spectrum in the absence of modulation) as narrow as 1 MHz. As will be seen, source linewidth is an important factor for certain noise-generating mechanisms.

The key characteristics of F–P and DFB lasers are covered next. Although this discussion applies generally to either type, the typical values given will be applicable to DFB units optimized for direct modulation, unless otherwise specified.

Linearity as a Directly Modulated Source

The typical transfer curve showing the relationship between driving current and light output power for a DFB is shown in Figure 12.14. In modern diodes that are designed for analog modulation, a C/CSO of 60 dB or better and a C/CTB of 65 dB or better are common for a signal loading of 80 unmodulated carriers and for signals that stay within the linear portion of the curve. The sharp discontinuity at low currents, however, causes abrupt clipping of applied waveforms that are too large. This is discussed later in this chapter.

Relative Intensity Noise (RIN)

The light from laser diodes has some random amplitude fluctuation. This is generally expressed in terms of the noise power in a 1-Hz bandwidth compared with the average optical power level. Typical values for quality DFBs are

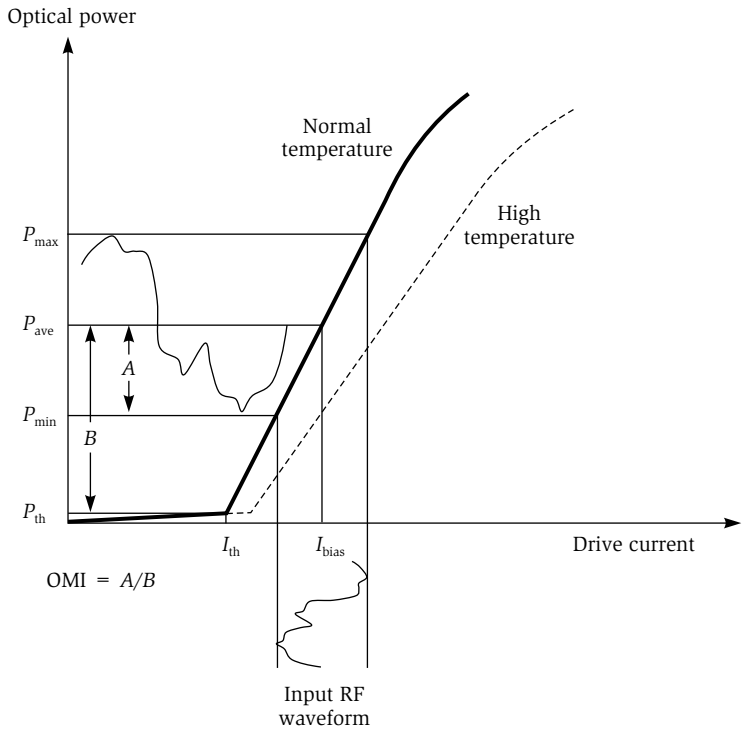


Figure 12.14 Laser diode transfer function.

approximately −160 dB/Hz, provided there is no reflected light. Any light reflected from the load, however, will degrade this quickly. For that reason, high-performance DFB sources are connected to fibers through optical isolators that serve to prevent reflected light from reaching the source.

In order to assess the effect of source RIN on link C/N, we need to compare it to the modulation of the light source by each carrier. The optical modulation index (OMI) is defined as the peak optical power variation from the unmodulated level divided by the the difference between the unmodulated level and the threshold level (A/B in Figure 12.14). Since the device is highly linear, OMI can also be defined as the peak deviation from the bias current level relative to the difference between bias and threshold current levels.

We can speak of the total modulation of the transmitter by the sum of all the RF carriers or of the modulation by some individual carrier. Though it is important that the total modulation not exceed 100%, to keep the signal from being clipped, we are also frequently interested in the performance of individual channels. The peak modulation of the optical carrier by an individual carrier is

abbreviated m_i, which will be an important factor in calculating the per-channel noise performance of an optical link.

The RIN contribution to total link C/N as a function of source RIN, m_i, and channel bandwidth is expressed by the formula

$$C/N_{RIN} = -\text{RIN} - 10\log(\text{BW}) + 20\log\left(\frac{m_i}{\sqrt{2}}\right) \tag{12.11}$$

where

C/N_{RIN} = the contribution of the source noise to the C/N of a signal, expressed in dB

RIN = the source noise level (relative to the unmodulated light power), expressed in dB/Hz

BW = the receiver noise bandwidth, in Hz, for the communications channel being evaluated

m_i = the peak modulation of the light source by the signal

As an example, if a source has an RIN of −160 dB/Hz and is modulated 3% (typical for a 77-channel system) by an NTSC video channel whose noise is measured in a 4-MHz bandwidth, the C/N will be approximately 60.5 dB. This puts a length-independent ceiling on the performance of a fiber-optic link (which will be further degraded by detector and postamplifier noise, as discussed later.)

Incidental Wavelength Modulation (Chirp)

When the current through a directly modulated laser diode changes as a result of modulation, there is also a small effect on the instantaneous wavelength of the produced light. This incidental FM is known as *chirp*. In typical laser diodes, the rate of frequency shift can vary over a range from approximately 50 to 500 MHz/mA with a total shift of ±1.5 to ±15 GHz, with DFBs exhibiting lower shifts than F–Ps and 1550-nm lasers exhibiting higher shifts than 1310-nm lasers.[13]

Chirp has several important effects on link performance. First, it spreads the energy over a range of wavelengths, which increases the effective Brillouin scattering threshold and thus allows higher optical power levels to be transmitted. On the other hand, it interacts with chromatic dispersion to cause a decrease in C/CSO. Finally, it interacts with reflected light to degrade effective source RIN.

Because the energy of a directly modulated DFB source is spread over a wider spectrum, it is possible to transmit higher power levels without reaching the Brillouin scattering threshold. Experience has shown that launch powers in excess of 25 mW into standard 1310-nm fiber are possible without a discernable increase in noise or distortion.

CSO Caused by Chirp Interacting with Dispersion

Over the relatively small wavelength shifts associated with chirp, the fiber's index of refraction will shift in an approximately linear manner. Thus, as the instantaneous modulating signal swings in one direction, the light will shift to a wavelength that travels slightly faster, whereas at the other polarity the generated light will travel slightly slower. Because the variance is approximately linear with modulation, the resultant distortion in the detected signal will affect one peak of the RF waveform exactly opposite of its effect on the other peak. As was discussed in Chapter 10, distortion with this characteristic causes only even-order products, of which the dominant is C/CSO. The relationship between chirp, RF frequency, fiber dispersion, and link length is given by the following formula:

$$C/CSO_{\text{CHIRP}-\text{D}} = -20\log(DL\ \Delta\lambda f_{RF} 2\pi 10^{-6}) - 10\log(N_2) + 3 \qquad (12.12)$$

where

$C/CSO_{\text{CHIRP}-\text{D}}$ = the ratio of the level of the visual carrier in question to the level of the CSO products falling in that channel

D = the fiber chromatic dispersion, in ps/nm-km, at the optical frequency used

L = the length of the fiber in km

$\Delta\lambda$ = the laser chirp per channel, in nm, as a result of modulation

f_{RF} = the frequency of the CSO product being analyzed, in MHz

N_2 = the number of second-order products in the composite beat that falls in the channel being analyzed (see Equation (10.25))[14]

The CSO affects channels differently because (1) a given amount of time shift represents more degrees of phase shift at higher frequencies, and (2) the number of second-order products falling into channels varies.

If the laser chirp per channel is known in terms of optical frequency rather than wavelength, it can be converted to wavelength chirp using

$$\Delta\lambda\ (\text{nm}) = 0.0057\ \Delta f\ (\text{GHz})\ @\ 1310\,\text{nm} = 0.0067\ \Delta f\ (\text{GHz})\ @\ 1550\,\text{nm} \qquad (12.13)$$

For example, typical 1310-nm DFBs exhibit about 100 MHz/mA incidental frequency modulation. If such a laser were modulated at 2 mA/channel and the signal transmitted through 20 km of standard fiber with a chromatic dispersion of 3 ps/nm-km, the chirp-related C/CSO at 550 MHz (31 upper-side second-order products) would be 60.6 dB.

Directly modulated 1550-nm DFBs tend to have higher chirp values — typically 250 MHz/mA. If such a source were modulated at the same 2 mA/channel and the light transmitted through the nonshifted 20-km fiber described earlier, the C/CSO would decrease to 35.8 dB due primarily due to the 18-ps/nm-km chromatic dispersion at that frequency. Thus, directly modulated DFB transmitters are only suitable for multichannel FDM video transmission when operated near the zero-dispersion wavelengths of fibers or when modulated by an RF spectrum of less than an octave so that the second-order products fall out of band.

Effects of Reflected Light

When light is reflected back into a DFB (or F–P) laser from a connector, passive component, receiver, or the fiber itself due to Raleigh backscattering, it can have a significant effect on RIN. Because the wavelength of the laser changes with modulation, it is almost certain that its wavelength will be different from that of the reflected signal due to the round-trip transmission time from transmitter to point of reflection.

If the laser wavelength is close to that of the reflected signal, it will try to lock its present oscillation to the frequency of the returned light. However, this is not the frequency the laser "should" be on, based on the modulating signal. The result is similar to that of an electronic oscillator that is receiving energy at a frequency close to, but not exactly on, its own frequency. The oscillator "pulls" toward the incoming frequency, resulting in a noisy output. The same thing happens with a laser and results in spurious emissions and an overall degradation in effective RIN. Some protection is afforded where optical isolators (components that exhibit low forward loss but typically greater than 30 dB of isolation in the reverse direction) are installed in laser diode output fibers, but attention to reflections from the driven optical circuits is also required. The jumps in frequency also add a degree of uncertainty to the calculation of interferometric intensity noise (see Section 12.9.1).

Temperature Effects

The current-versus-light transfer function for diodes shown in Figure 12.13 is stable at any one temperature. As the temperature changes, the curve shifts horizontally and may also change slope. Since this could result in a shift of the operating point from the optimum value, lasers used for wideband, downstream transmission are almost always connected to thermoelectric coolers (TECs) that act to maintain a constant chip temperature.

While the use of TECs would be desirable in all linear modulation applications, the cost and power consumption have resulted in the use of noncooled transmitters in many lower-performance upstream transmitters.

12.6.4 Externally Modulated Continuous Wave Sources

Externally modulated transmitters consist of a continuous wave light source whose intensity is varied through the use of an external device that is driven by the FDM waveform. This configuration offers several advantages, including:

- Freedom from the incidental wavelength modulation (chirp) associated with direct modulation of laser diodes (though modulators have some residual shift of their own)
- Improved isolation of the source from reflections in the transmission system, thereby decreasing effective source RIN
- The ability to use a variety of optical sources without regard to their ability to vary light output in proportion to an electrical modulating signal

The most common external modulator is a Mach–Zehnder (M–Z) device constructed on a lithium niobate ($LiNbO^3$) substrate. In an M–Z modulator, the incoming light is divided proportionately between two outputs in response to an electrical stimulus. The transfer function to each output is very nearly sinusoidal, as illustrated in Figure 12.15. When used as a linear transmitter for cable television distribution, the predictable symmetric curvature of the transfer function (which causes large odd-order distortion products) is compensated for by predistortion in the driver circuit.

The transfer function of M–Z modulators provides one important additional advantage over directly modulated sources: The distortion is symmetrical about the inflection point of the transfer function, offering significantly improved

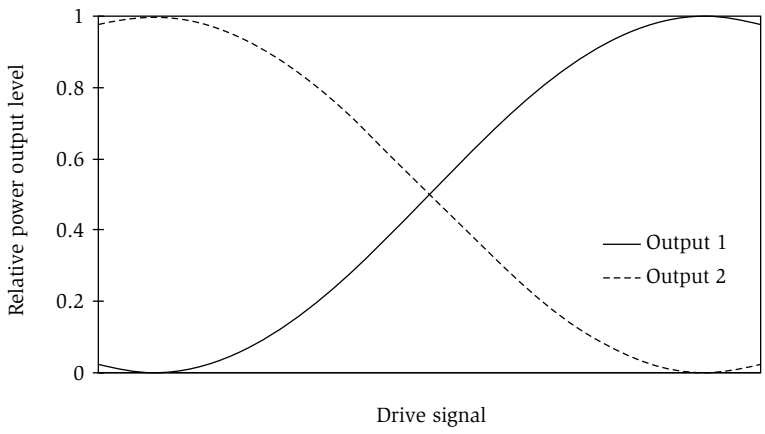


Figure 12.15 Transfer function, external optical modulator.

second-order distortion products (see Chapter 10, dealing with distortion in coaxial systems).

Externally modulated transmitters can be constructed at either 1310 or 1550 nm. At 1310 nm, the usual source is a relatively high-powered (80-mW typical) YAG oscillator. This device, whose wavelength is actually 1319 nm, exhibits a low RIN and optical linewidth.

At 1550 nm, it is more common to use a DFB diode for the light source and then to amplify to the desired power level using optical amplifiers. Because an unmodulated DFB has a spectral width of less than 100 MHz, this makes Brillouin scattering a concern.

Regardless of the light source, when an FDM spectrum extending up to between 500 MHz and 1.0 GHz is used to amplitude modulate the light, the optical spectrum will have sidebands extending to the maximum modulating frequency on each side of the unmodulated optical center frequency. It might be expected that this would effectively raise the Brillouin limit. An analysis, however, has shown that 99% of the energy is still concentrated at the unmodulated wavelength, and actual measurements have demonstrated a negligible increase in scattering limit.[15]

High launch powers are possible, however, if the total optical power is spread more evenly over a range of wavelengths. Two methods are commonly used: phase modulation in the same external modulator used for AM modulation and, in the case of DFB sources, intentional chirp generated by modulating the DFB's operating point with a constant frequency. In either case, the wavelength spreading must be sufficient that the beats between the optical frequencies fall above the highest modulating frequency. Manufacturers who employ SBS extension techniques commonly claim the ability to launch +16 to +17 dBm into long single mode non-dispersion-shifted fibers with acceptably low levels of CSO degradation at either 1310 or 1550 nm.

12.7 Optical Amplifiers

Optical amplifiers boost the level of amplitude modulated light without requiring the optical-to-electrical conversion, RF amplification, and remodulating otherwise required in a repeating location in a network. Understanding the process requires a knowledge of atomic physics that is beyond the scope of this book, but a broad explanation of the principles follows.

The most common form of amplifier is the 1550-nm erbium-doped fiber amplifier (EDFA). In an EDFA, the incoming signals are combined, in a WDM coupler, with a high-power unmodulated signal at 980 nm* from one or more

* 980 nm is the most commonly used pump wavelength. Older devices used 1480-nm pumps, and other wavelengths, such as 850 nm, are possible.

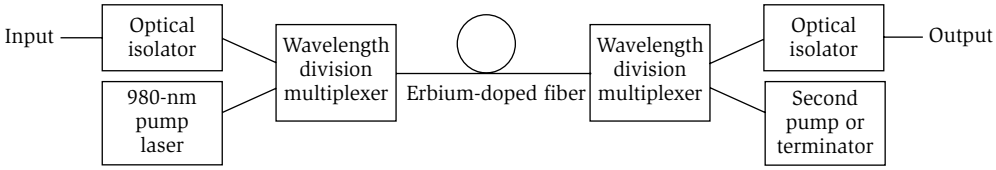


Figure 12.16 Schematic of EDFA.

"pump" sources. The two signals are then sent through a special length of fiber that is "doped" with erbium (that is, a carefully controlled amount of erbium is added). The pump signal causes electrons in the erbium atoms to jump to a higher energy band. When they return to their normal state, the difference in energy boosts the signals at 1550 nm. Figure 12.16 shows one of several possible basic schematics of a dual-pump EDFA.

The magic of EDFAs is that they have completely different characteristics than coaxial amplifiers with respect to noise and distortion. Even though they operate in a saturated mode (where a normal coaxial amplifier would produce pulses with a high harmonic and IM content), EDFAs add no measurable distortion to the modulated signal. Not only that, but they can be overdriven, if desired, so that their input noise level is small compared with the signal to be amplified. The C/N contribution from an EDFA is given by

$$C/N_{EDFA} = 86.2 + P_i + 20\log(m_i) - \mathrm{NF}_{EDFA} \tag{12.14}$$

where

C/N_{EDFA} = the carrier-to-noise per channel (measured in a 4-MHz bandwidth) in dB

P_i = the optical input power to the EDFA in dBm

m_i = the optical modulation index (OMI) per carrier

NF_{EDFA} = the noise Figure of the amplifier

The noise Figure of a standard EDFA depends somewhat on input power. Figure 12.17 shows typical variations in NF based on several manufacturers' specifications.

Commercial amplifiers are available with saturated power output levels in excess of 100 mW. The highest-powered units utilize two or more pump sources, injecting light into ends of the doped fiber (actually, four pumps are possible, two at each end, if they are cross polarized), while multiple-stage devices are also common. Increasing pump power increases the available output power

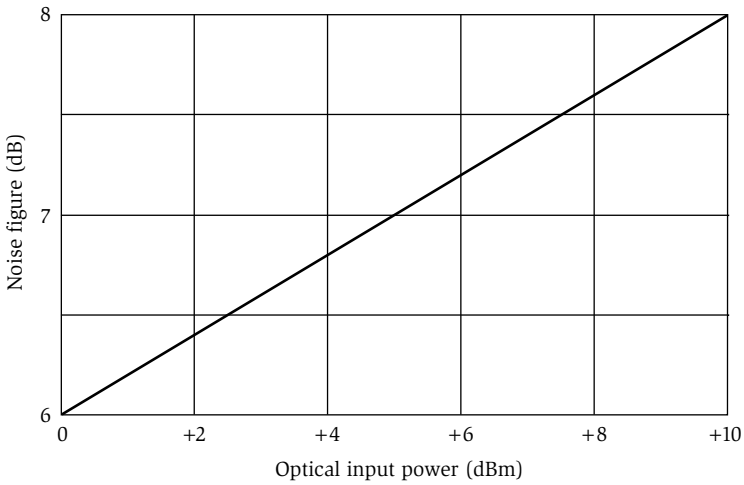


Figure 12.17 Noise figure variation in EDFA as a function of drive level.

since more energy is stored in excited electrons that can be tapped to produce optical emissions. A modified technology uses erbium and ytterbium doping and an 800-nm pump wavelength and achieves even higher powers, potentially in the hundreds of milliwatts.[16]

Unfortunately, the efficiency and performance of an optical amplifier are accidents of the properties of specific atomic elements. Although 1550-nm amplifiers are a mature technology, it has proved much more difficult to produce commercially practical units at 1310 nm. Recently reported results, though, include a +20-dBm unit with a 22-dB small signal gain and 6-dB noise Figure. This unit uses praseodymium doping.[17]

12.8 Optical Receivers

Unlike transmitters, optical detectors are relatively standardized. The incoming light impinges on the active area of a photodiode that has a very precise square law transfer curve (that is, the output current is proportional to input optical power). Typical responsivity (the slope of the received optical power versus output current transfer function) is 0.8–1.0 mA/mW.

The noise performance of an optical receiver is limited by the noise current of the diode (shot noise due to the statistical variation in arriving photon distribution) and by the noise of the following coaxial amplifier. Since the inherent impedance of the detector diode is not 75 ohms, either an impedance-matching transformer or an active circuit is required to match it to the gain stages. Some-

times a *transimpedance amplifier* is used. Though these amplifiers have an excellent noise figure, they have limited dynamic range, with the result that the light level reaching a receiver must typically be controlled within a few decibels of 0 dBm for the best balance between noise and distortion for AM modulated links. It is common for manufacturers of analog broadband fiber-optic terminal equipment to specify link performance at or near 0-dBm received power.

The formula for the C/N contribution due to shot noise is

$$C/N_{\text{SHOT}} = P_r + 20\log\left(\frac{m_i}{\sqrt{2}}\right) + 10\log(R) - 10\log(\text{BW}) + 154.94 \qquad (12.15)$$

where

C/N_{SHOT} = the C/N of an individual carrier due to shot noise in the detector, expressed in dB

P_r = the received optical power level in dBm

m_i = the peak modulation of the light source by the carrier

R = the responsivity of the receiving diode in amperes per watt (or mA/mW)

BW = the noise susceptibility bandwidth of the channel in Hz

For example, if the OMI per carrier is 3%, the diode response is 0.9 A/W, the received optical power is 0 dBm, and the video bandwidth is 4 MHz, then the noise contribution due to shot noise will be approximately 55 dB.

Added to that must be the contribution from the postdetector amplification system. If a transimpedance amplifier is used, its C/N contribution can be calculated using

$$C/N_{\text{POSTAMPLIFIER}} = 2P_r + 20\log\left(\frac{m_i}{\sqrt{2}}\right) + 20\log(R) - 10\log(\text{BW}) + 10\log(R_Z) - F + 137.91 \qquad (12.16)$$

where

$C/N_{\text{POSTAMPLIFIER}}$ = the C/N of a signal due to thermal noise generated within the postamplifier, expressed in dB

R_Z = the postamplifier transimpedance in ohms

F = the postamplifier noise Figure in dB

Note that the C/N varies 2:1 with received optical power (the factor $2P_r$) because of the square law nature of the detector, as opposed to C/N_{SHOT}, which varies 1:1 with optical power because it occurs in the optical domain (before detection).

If, in the previous example, the detector diode is attached to a postamplifier having a 3-dB noise figure and 1200-ohm transimpedance, the noise contribution due to postamplifier noise will be 65.3 dB. Transimpedance amplifiers, however, have very limited dynamic range. In commercial optical receivers optimized for cable television applications, a variety of means are used to couple photodiodes to postamplifiers, including both direct and impedance-matching transformers. A more general way to calculate postamplifier noise uses the total postamplifier equivalent thermal noise (which includes both the irreducible noise due to the impedance and the excess noise generated within the amplifier).

If Equation (12.16) is reconfigured to use equivalent amplifier thermal noise input current, the postamplifier C/N contribution can be calculated using[18]

$$C/N_{\text{POSTAMPLIFIER}} = 2P_r + 20\log\left(\frac{m_i}{\sqrt{2}}\right) + 20\log(R) - 10\log(\text{BW}) - 20\log(I_r) + 180 \tag{12.17}$$

where

$C/N_{\text{POSTAMPLIFIER}}$ = the C/N of a signal due to thermal noise generated within the postamplifier, expressed in dB

P_r = the received optical power level in dBm

m_i = the peak modulation of the light source by the signal being evaluated

R = the responsivity of the photodiode in amperes/watt (or mA/mW)

I_r = the postamplifier equivalent input noise current density in $\text{pA}/\sqrt{\text{Hz}}$

180 = a factor that results from converting the input power to milliwatts from watts and the amplifier noise to picoamperes from amperes

For example, if we use the same parameters as in the last example but assume instead that the amplifier input noise current density is $7\ \text{pA}/\sqrt{\text{Hz}}$ (typical values will be 6–8), we can calculate that the postamplifier C/N for an analog video channel will be 62.7 dB.

12.9 Interactions Among Transmitters, Fibers, and Receivers

In addition to the noise and distortion that take place within transmitters and receivers, additional signal degradation takes place because of the interaction of fiber characteristics with the characteristics of the optical transmitter and receiver.

These additional effects can have a significant effect on link performance, which is why it is important that fiber-optic equipment be specified and tested through actual fiber links, not just attenuators.

12.9.1 Double Rayleigh Backscattering: Interferometric Intensity Noise

Some portion of light that is scattered backward through the fiber due to Rayleigh scattering will be rescattered in the forward direction and will combine with the normally transmitted light. Since the sum of the double-scattered light from various lengths will not be coherent with the normal light, the effect when both reach the optical detector is that they mix to produce products at the differences in optical frequencies (and multiples thereof) and will show up as an effective increase in link noise. This interferometric intensity noise (IIN) is spread across a spectrum that is about twice the total effective linewidth of the transmitter (including the effect of chirp, if present). The effective link IIN is given by

$$\mathrm{IIN} = 10\log\left[\frac{3.6\times10^{-14}\left[L-\frac{1}{2\alpha}\left(1-e^{-2\alpha L}\right)\right]}{\Delta f_{RMS}}\right] \tag{12.18}$$

where

$\alpha = 1-10^{-\frac{\alpha_0}{10}}$

IIN = the level of double-scattering-caused noise relative to the unmodulated light level, in dB/Hz

L = the length of the fiber in km

α_0 = the loss of the fiber in dB/km

Δf_{RMS} = the total rms effective linewidth (in the case of directly modulated transmitters, this is the total wavelength spread, chirp, of the source as a result of modulation; in the case of externally modulated transmitters, this is the linewidth of the unmodulated source if not dithered or phase modulated)[19]

For affected frequencies, IIN can be related to in-channel C/N in the same way as transmitter RIN:

$$C/N_{IIN} = -\mathrm{IIN} - 10\log(\mathrm{BW}) + 20\log\left(\frac{m_i}{\sqrt{2}}\right) \tag{12.19}$$

where

C/N_{IIN} = the contribution of the IIN effects to the C/N of a signal, expressed in dB

IIN = the interferometric noise level (relative to the unmodulated light power), as calculated in Equation 12.18 and expressed in dB/Hz

BW = the receiver noise bandwidth, in Hz, for the communications channel being evaluated

m_i = the peak modulation of the light source by the signal

Figure 12.18 shows an example of the effect of IIN on video C/N as a function of fiber length for several optical linewidths at 1310 nm. The assumed OMI per carrier was 3%.

For externally modulated sources, absent SBS suppression techniques that cause considerable line spreading, the linewidth is so narrow that IIN noise density is very high and is concentrated at very low frequencies. In order to avoid severe IIN degradation to television channel 2, which starts at 54 MHz, the linewidth must be either less than a few megahertz or artificially spread to levels at least comparable with those found with directly modulated lasers. Since a wider spectral line may be required as part of SBS suppression techniques, the latter approach is generally taken.

Finally, IIN may affect long upstream DFB links in a different way. The unmodulated linewidth of a typical DFB is approximately 10 MHz. Since an

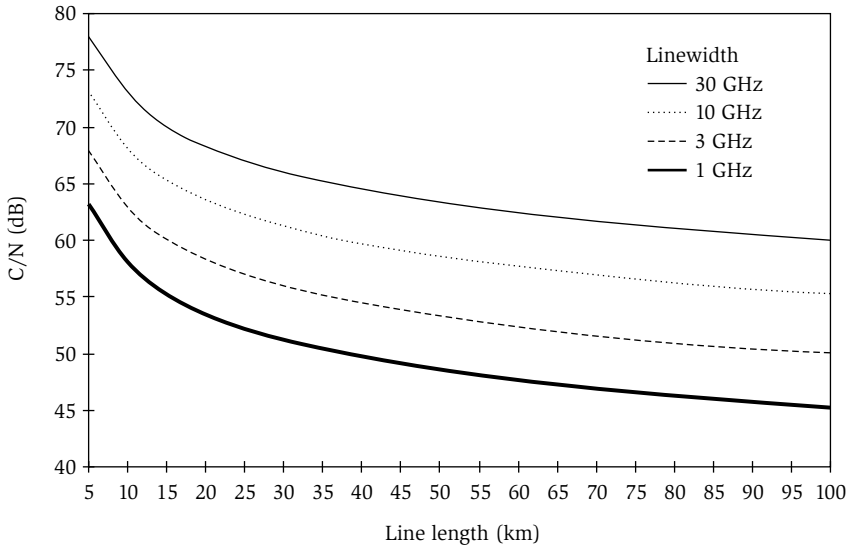


Figure 12.18 C/N degradation due to IIN.

upstream link may not be used continuously, when it is idle there may be a significant level of IIN extending up to about 20 MHz much of the time. A possible answer is to apply a "dithering" modulation to the laser continuously to ensure a wider chirped linewidth.[20]

12.9.2 Phase Noise Contribution to Link Performance

Even though DFB diodes oscillate in only a single longitudinal mode, they still exhibit some residual frequency instability. As in F–P diodes, this variance interacts with dispersion to create link noise.

The relationship between source phase noise and link C/N is given by the formula[21]

$$C/N_{PM-AM} = 398.55 - 10\log(\Delta v\ \mathrm{BW}_{RF}\ D^2\lambda^4 f_{RF}^2 L^2) + 20\log\left(\frac{m_i}{\sqrt{2}}\right) \qquad (12.20)$$

where

C/N_{PM-AM} = the contribution to channel C/N (in dB) due to the phase noise in the optical source being converted to amplitude noise in the detector

Δv = the source linewidth in MHz

BW_{RF} = the noise susceptibility bandwidth of the channel in Hz

D = the fiber dispersion constant in ps/nm-km

λ = the operating free-space optical wavelength in nm

f_{RF} = the modulating frequency in MHz

L = the fiber length in km

m_i = the OMI per RF carrier at f_{RF}

The 398.55 factor is the result of converting all other values to common engineering units.

As can be seen, the noise affects the RF channels unequally, increasing as the square of the RF channel frequency. It also increases as the square of the fiber length.

Phase-related amplitude noise is generally not of concern in 1310-nm links using nonshifted fiber but may be significant in cases of long fiber links using non-dispersion-shifted fiber at 1550 nm. As an example, assume a 750-MHz analog video channel used to externally modulate (OMI/ch = 3%) an amplified 1550-nm DFB whose optical linewidth is 1 MHz. The signal is then transmitted through a 60-km fiber whose chromatic dispersion is 17 ps/nm-km. The detected signal will exhibit a C/N due to PM–AM conversion of 53.8 dB.

12.10 End-to-End Fiber-Optic Link Performance

The characteristics of the transmitter, interconnecting fiber, receiver, and, optionally, any amplifiers work together to determine the link performance. In particular, the interaction between the signals and the transmission media is much more complex in optical than in coaxial links.

12.10.1 Noise Performance

The total link C/N will have contributions due to transmitter RIN, detector shot noise, and postamplifier noise Figure as well as due to the IIN interaction among laser linewidth, double scattering, and detector mixing. These noise effects are noncorrelated and so will add on a power basis:

$$C/N_{\text{LINK}} = -10\log\left[10^{-\left(\frac{C/N_{RIN}}{10}\right)} + 10^{-\left(\frac{C/N_{IIN}}{10}\right)} + 10^{-\left(\frac{C/N_{\text{SHOT}}}{10}\right)} + 10^{-\left(\frac{C/N_{\text{POSTAMPLIFIER}}}{10}\right)}\right] \tag{12.21}$$

where

C/N_{LINK} = the net per-channel C/N

C/N_{RIN} = the per-channel transmitter relative intensity noise (Equation (12.11))

C/N_{IIN} = the per-channel interferometric intensity noise (Equation (12.19))

C/N_{SHOT} = the per-channel shot noise at the detector input (Equation (12.15))

$C/N_{\text{POSTAMPLIFIER}}$ = the per-channel postamplifier noise contribution (Equation (12.16) or (12.17))

If the end-to-end C/N of a simple DFB link is plotted as a function of optical loss, the curve will typically have three asymptotes, as shown in Figure 12.19 (the effects of IIN are not readily plotted on the same chart because they vary not with received optical power but with line length).

For high received power levels, the transmitter RIN will dominate, and the C/N will be independent of path loss; at lower received levels, the shot noise will begin to dominate, and the link C/N will change 1 dB for every decibel of reduction in received power level. At some point, the postamplifier noise will begin to dominate, and the link will begin to degrade 2 dB for every decibel of loss, due to the square law receiver response. The range of receiver power over which each of these effects dominates depends, of course, on the quality of the

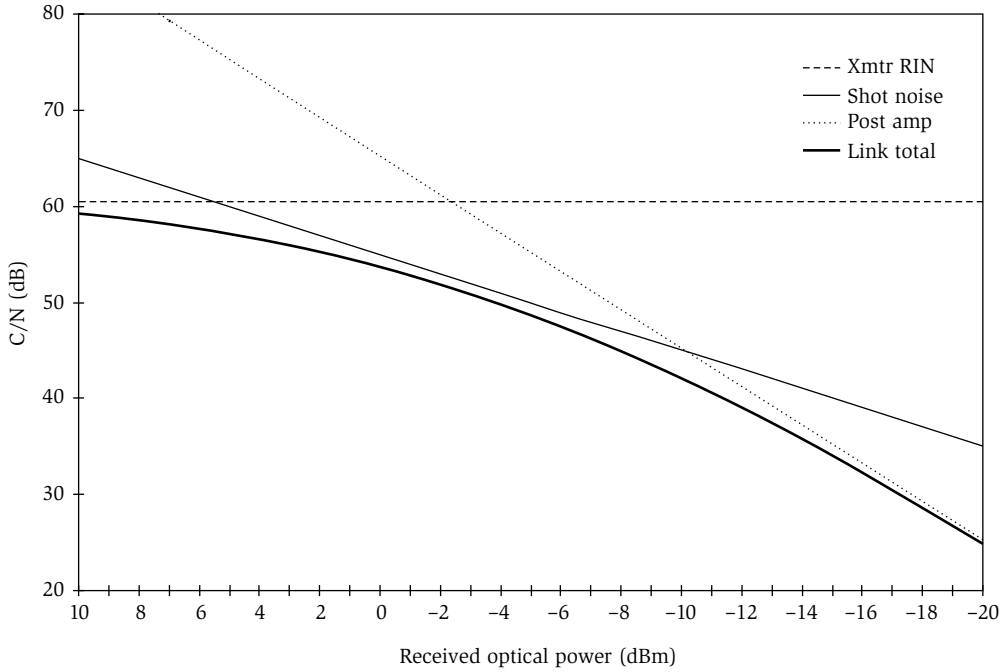


Figure 12.19 Typical fiber-optic link C/N contributions.

individual components. For typical links, the slope of C/N versus level near 0-dBm received power is about 1:1.

In addition to the factors shown in Equation (12.21), other elements and interactions may affect the end-to-end C/N. For instance, if an EDFA is used, its C/N must be included within the parentheses as an additional factor. In long 1550-nm links, residual phase noise may be a factor.

12.10.2 Small-Signal Distortions

Directly modulated DFB transmitters behave similarly to single-ended RF amplifiers, generating both second- and third-order distortion whose amplitudes, as a function of RF drive level, increase roughly 1 dB/dB (in the case of second-order products) and 2 dB/dB (in the case of third-order products). As with their coaxial counterparts, CSO products tend to be the limiting distortion. As discussed earlier, typical specifications for such transmitters are 60–62 dB for C/CSO and 65 dB for C/CTB under normal operating conditions with a loading of 77 unmodulated carriers. These distortions arise from the combination of small transfer function nonlinearities and large signal clipping.

Externally modulated transmitters exhibit symmetrical distortion and are thus similar to conventional push–pull coaxial amplifiers in that regard. That, combined with the predictability of the nonlinearities, allows C/CSO and C/CTB values of 65 or better with channel loadings of 77 unmodulated carriers.

Distortion occurring in the optical receiver must be added to that generated in the transmitter. Like the transmitter, the detector–postamplifier combination behaves like a single-ended amplifier, with dominant second-order distortion. Often manufacturers specify link performance with a given received power, rather than individually specifying transmitter and receiver.

When the link loss includes fiber, the CSO degradations due to both chirp-induced dispersion and SPM must be included in the calculation of end-to-end performance. Since the CSO distortion mechanisms are synchronized primarily with the modulating waveform (with the exception of the small-signal DFB nonlinearities), the various CSO contributions could be expected to add on a voltage (20 log) basis rather than on a power basis. Though this is true, the chirp and SPM effects may, in fact, be of opposite polarity and tend to cancel each other, rather than add. Nevertheless, an assumption of voltage addition is the most conservative approach to link design.

One technique for reducing the effects of chirp, SPM, and receiver-generated CSO is to transmit both outputs of a Mach–Zehnder type modulator through parallel networks to the receiver. If the two signals are separately detected and then combined out of phase at RF, these second-order effects will tend to cancel, just as they do in push–pull RF amplifiers. C/CSO values of greater than 70 dB have been reported using this technique, known commercially as a Harmonic Link Extender®. As an added benefit, when the signals are combined, the noise will be largely uncorrelated (except for the contribution from transmitter RIN) and so will add on a power basis, whereas the signals will add on a voltage basis, resulting in a net 3-dB increase in C/N in the link. Link performance of 55-dB C/N over 50-km links at 1550 nm has been reported.[22]

12.10.3 Clipping Distortion

The ratio of peak-to-average absolute voltage in a simple unmodulated carrier is about 1.6:1. When two equal-amplitude nonrelated carriers are summed, however, the relationship is not as simple. Figure 12.20 shows three possible waveforms resulting from the simple addition of two sine waves when one is exactly twice the frequency of the other. Depending on their relative phases, the peak voltages can vary by nearly a factor of 2.

The case when many signals are summed is much more complex. When many summed signals are randomly distributed in both frequency and phase, the peak-to-average level seldom exceeds 14 dB, as discussed in Chapter 15 (though with occasional higher peaks). The downstream spectrum of cable systems is not random, however. The most extreme case occurs when a large

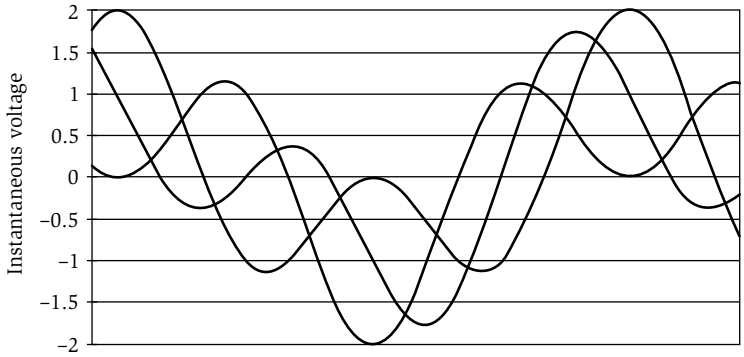


Figure 12.20 Addition of two sine waves with various phase differences.

number of carriers are harmonics of a common root frequency (as in the ANSI/EIA-542 HRC channelization plan) and all the carrier phases are aligned. In that case, the time domain waveform can resemble a string of impulses spaced by a time interval equal to the period of the common root frequency. This is not a surprising result since the Fourier transform of an infinite string of zero-width pulses is an infinite string of carriers.[23]

In the more commonly used EIA-542 Standard frequency plan, most of the visual carriers are offset by approximately 1.25 MHz from harmonics of 6 MHz and are neither frequency nor phase related. Despite this, the composite waveform has characteristics that are close to harmonically related carriers if the carrier phases do happen to align at some time. Figure 12.21 shows the resultant instantaneous voltage of the standard channels between 54 and 550 MHz, with each assigned a frequency at random within its tolerance range but with the signal phases aligned at $t = 30$ ns on the horizontal scale. The peak-to-average voltage is approximately 34:1. If we look at the same waveform later in time, however, the phases slowly move out of coincidence, and the peak voltage decreases. Figure 12.22 shows the condition at $t = 15$ μs. By 20−30 μs, the pulses are no longer distinguishable. Figure 12.23 shows a typical situation with truly random phase relationships.

Observations of actual headends carrying NTSC modulated carriers have confirmed that peak-to-average ratios generally vary from about 3.5:1 to greater than 7:1. When some subset of carriers happen to phase align, the result is voltage peaks spaced approximately 167 ns apart (the inverse of the 6-MHz channel spacing), which build up to a peak over a few tens of microseconds and then decrease again as the carriers move out of phase coincidence.

The importance of this occasionally occurring peak voltage condition is that, depending on the OMI, the optical transmitter may be driven into hard limiting (clipping) when a sufficient number of carriers are in phase alignment. This is

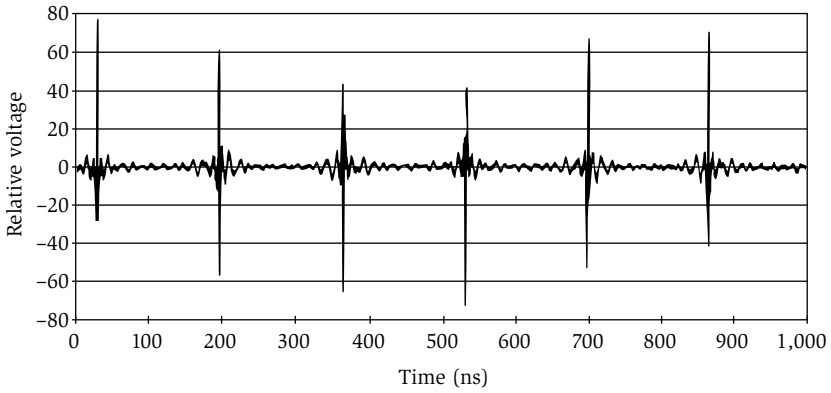


Figure 12.21 Instantaneous voltage from addition of 77 FDM carriers with standard frequency assignments and tolerances with phases aligned at $t = 30$ ns.

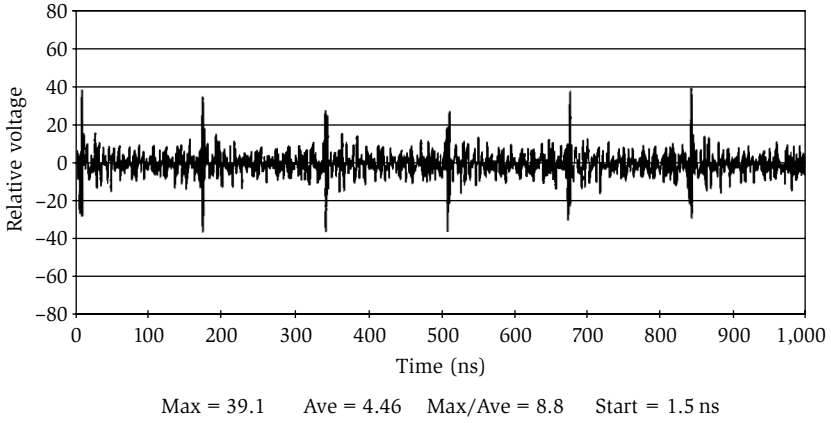


Figure 12.22 Spike degeneration after 15 microseconds.

particularly true in the case of directly modulated laser diodes, where a sharp knee occurs in the transfer function below which the light output is extinguished. When that occurs, intermodulation products are generated. As Figure 12.24 shows, both even- and odd-order distortion products are generated when one polarity of a sine wave is clipped off (though second-order products are always the largest). Thus, both CTB and CSO will show an increase when

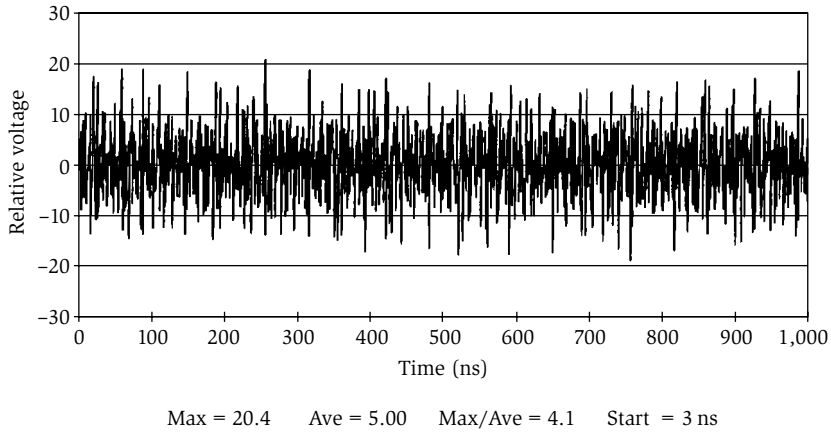


Figure 12.23 Typical headend output waveform.

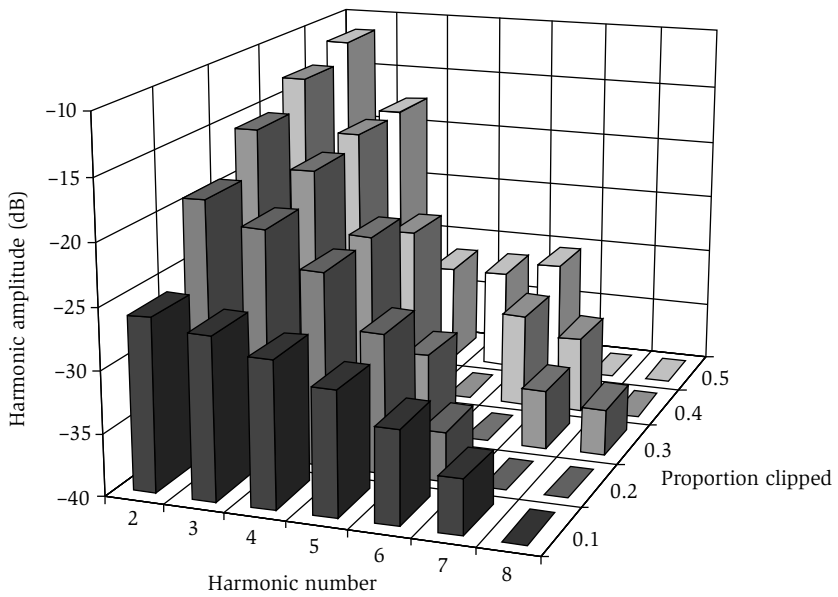


Figure 12.24 Harmonic content of clipped sine wave.

clipping happens with sufficient frequency to be statistically significant. Since clipping produces a string of pulses with the characteristic 167-ns spacing, the largest components produced will fall at harmonics of 6 MHz (which is at the position of the lower CSO beat when the EIA Standard or IRC channel allocation scheme is used and at the video carrier frequency when the EIA HRC channel scheme is used). With highly linear diodes, clipping may, in fact, be the principal contributor to IM products.

In the case of Mach–Zehnder external modulators, the nonlinearities are symmetrical, and the limiting is "softer," resulting in less clipping and primarily odd-order resultant distortion.[24]

The visible effect on analog video channels is that, as OMI is increased, "dashes" resembling electrical interference begin to appear on the displayed picture. These visible effects are due to the integrated effect of a group of pulses (whose spacing is closer than the television can resolve).

In the case of high-order digitally modulated carriers sharing a linear link with a comb of analog television carriers, the result of clipping is an increase in bit error rate (BER). Data taken by one of the authors found that the BER through a fiber link loaded with 80 normally modulated video channels and one 64-QAM data channel (whose signal level was 10 dB below the video channels) was approximately 1.3×10^{-10}. As the OMI was increased, the BER did not change until a critical threshold was reached, whereupon the BER increased to 7.6×10^{-8}, a 600:1 degradation, with an OMI increase of less than 1 dB. This confirms findings of other researchers in this field. The observations of researchers using normal NTSC video signals were that the BER increased to unacceptably high values before any degradation was visible in displayed analog pictures.[25]

The crucial question is how often clipping occurs. This has been explored by a number of authors. David Grubb has suggested that if OMI per channel is held to $0.348/\sqrt{N}$ (where N is the number of channels), the voltage will exceed the clipping threshold about 2.5×10^{-5} of the time.[26] That analysis, however, is based on unmodulated, randomly phased carriers. He noted, however, that optimal phasing of harmonically related carriers can reduce the theoretical peak voltage. The ratio of theoretical maximum voltage with phases aligned to the maximum achievable by optimum phasing is $(\sqrt{N})/1.5$. If the signals feeding an optical transmitter were so related, the modulation index could be increased to $0.67/\sqrt{N}$, or about 5.7 dB higher. This would translate directly to an achievable C/N increase in the link (whose exact amount is related to the relative contribution of transmitter, shot, and postamplifier noise, as discussed earlier in this chapter).

Another factor is video modulation. Since NTSC signals are at maximum carrier level only a small percentage of the time (during synchronizing peaks), the sum of the instantaneous voltages of a number of randomly timed video channels is almost always lower than unmodulated carriers. As with carrier phasing, however, the key word is *almost* since it is possible that the synchronizing pulses will occasionally drift into alignment. Stuart Wagner has measured

clipping noise and its effect on the BER of a digital signal carried along with 40 randomly timed analog video channels in a fiber link. What he found was that the peak OMI per analog carrier could be increased by approximately half the average difference in power level (in decibels) between the unmodulated and modulated carriers with no increase in noise or BER.[27] As discussed in Chapter 10, the usual assumption is that average power levels are about 6 dB below peak levels. In particular, the black level of a video signal is about 2.5 dB below sync, so with random video timing it should be possible, on average, to raise peak OMI per channel by about 3 dB compared with the unmodulated carrier case.

On the other hand, with optimum video timing it should be possible to keep the total carrier power close to the average power by minimizing instances where carriers are simultaneously at peak levels. In that case, it should be possible to increase OMI by the difference between peak and average power, 6 dB, without increasing clipping distortion.[28] That increase will lead to a similar increase in link C/N.

12.11 Summary

Linear fiber-optic links are capable of transporting the full spectrum of cable television services over distances exceeding 20 miles without amplification. Amplification, readily available at 1550-nm wavelength, can extend the attainable distance severalfold, with only a minor impact on end-to-end performance.

The basic C/N performance of optical links is limited by transmitter intensity noise, optical shot noise, the interaction between transmitter linewidth and doubly scattered light in the fiber, and receiver postamplifier noise. The interaction between transmitter residual phase noise and the fiber can further reduce the noise under some conditions.

Distortion is fundamentally limited by both small-signal nonlinearities in the transmitter and clipping caused by large-signal peaks. Additionally, interaction between incidental transmitter FM and fiber dispersion can cause second-order distortion, as can interaction between peak power levels and the glass material itself.

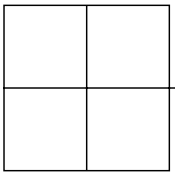
Increasing the optical modulation index per carrier in an optical link will improve C/N at the expense of distortion, just as increasing operating levels in a coaxial network will have the same effect. Similarly, decreasing the number of carriers will allow higher C/N and lower distortion per channel, as in a coaxial network.

Chapter 13 will discuss multiwavelength transmission through shared fibers. Chapter 15 will discuss the total end-to-end performance of cascaded fiber-optic and coaxial broadband distribution networks, Chapter 18 will discuss various HFC architectures, Chapter 19 will discuss fiber-deep achitecture, and Chapter 20 will discuss the reliability of various HFC architectures.

Endnotes

1. Ronald C. Cotten, *Lightwave Transmission Applications*. CableLabs, Louisville, CO, September, 1993, pp. 80–84.
2. Data provided by CommScope, Inc., Hickory, NC.
3. *AT&T Generic Specification: Fiber-Optic Outside Plant Cable*, Issue 11, December, 1995. AT&T Network Cable Systems, Norcross, GA.
4. *CommScope Fiber Optic Cable Catalog*, September, 1995. CommScope, Inc., Hickory, NC.
5. AT&T, op. cit.
6. Claude Romans and David Large, Optical Bus Architecture for Co-Deployment of Telephone and CATV Services in the FRG, *Conference Record, Globecom '89*. IEEE Communications Society, November 1989, pp. 37.5.1–37.5.9.
7. AT&T, op. cit.
8. David Grubb III and Yudhi Trisno, AM Fiber Optic Trunks — A Noise and Distortion Analysis, *1989 NCTA Technical Papers*, NCTA, Washington, DC, May 1989.
9. F.W. Willems, W. Muys, and J.S. Leong, Simultaneous Suppression of Stimulated Brillouin Scattering and Interferometric Noise in Externally Modulated Lightwave AM-SCM Systems, *IEEE Photonics Technology Letters*. Vol. 6, No. 12, December 1994. See also S. W. Merritt, G. J. McBrien, and E. R. Yates, Integrated Optic Modulators for 1-GHz HFC Systems, *CED*, February 1996.
10. Dogan A. Atlas, Fiber-Induced Distortion and Phase Noise to Intensity Noise Conversion in Externally Modulated CATV Systems, *1996 NCTA Technical Papers*, NCTA, Washington, DC.
11. AT&T Corporation, *LGX Broadband Fiber Management System Application Guide*, AT&T Network Systems, Norcross, GA, 1995.
12. Dan Harris, Primer on Baseband Digital Transmission on Single-Mode. *Communications Technology*, August 1996, pp. 92–98.
13. Ronald C. Cotten, op. cit., p. 148.
14. Henry A. Blauvelt et al., Optimum Laser Chirp Range for AM Video Transmission, *1992 NCTA Technical Papers*. NCTA, Washington, DC, pp. 79–85.
15. X. P. Mao et al., Brillouin Scattering in Lightwave AM-VSB CATV Transmission Systems. Paper presented at OFC '92 Conference, San Jose, CA, February 1992. Digested paper in *Technical Digest — Optical Fiber Communication Conference*. Published by Optical Society of America, Washington, DC, 1992.
16. Roger Brown, Wave of New Products Set to Break. *CED*, September 1996, pp. 26–30.

17. Vincent Morin and Edouard Taufflieb, Improved Optical Fiber Amplifier for 1.3-μm AM-VSB CATV Systems. Scheduled for presentation at the 1998 NCTA National Conference and publication in the *1998 NCTA Technical Papers*. NCTA, Washington, DC.
18. This equation is identical to that used by Ronald Cotten in the previously referenced paper (his Equation 16, page 171), but with the input variables converted to more common engineering units and with the bandwidth kept as a separate factor to allow the formula to be applied to various bandwidth services.
19. Henry A. Blauvelt, op. cit.
20. Donald Raskin and Dean Stoneback, *Return Systems for Hybrid Fiber/Coax Cable TV Networks*. Upper Saddle River, NJ: Prentice-Hall, 1998, pp. 152–154.
21. Dogan A. Atlas, op. cit. Modified after consultation with Dr. Dan Harris of Corning Fiber to give the result per channel and to convert the units to engineering units more familiar to cable engineers.
22. Patrick Harshman, Applying Push–Pull Technology to Optical Networks. *CED*, September 1996, pp. 56–61.
23. Athanasios Papoulis, *Circuits and Systems: A Modern Approach*. New York: Holt, Rinehart & Winston, 1980, pp. 331–389. Fourier transforms are used to convert between time domain waveforms and their equivalent frequency content.
24. D. Raskin et al., Don't Get Clipped on the Information Highway, *1996 NCTA Technical Papers*. NCTA, Washington, DC.
25. David Large and Rex Bullinger, Downstream Laser Clipping: Field Measurements and Operational Recommendations, *Proceedings Manual, 1997 Conference on Emerging Technologies*. SCTE, Washington, DC, January 1997.
26. David Grubb III and Yudhi Trisno. AM Fiber-Optic Trunks — A Noise and Distortion Analysis, *1989 NCTA Technical Papers*. NCTA, Washington, DC, 1989.
27. Stuart S. Wagner, Thomas E. Chapuran, and Ronald C. Menendez, The Effect of Analog Video Modulation on Laser Clipping Noise in Optical Video-Distribution Networks, *IEEE Photonics Technology Letters*. Vol. 8, No. 2, February 1996, pp. 275–277.
28. David Large, Reducing Distortions Using Video Timing Techniques. *CED*, June 1989, pp. 114–119.



Chapter 13
Wavelength-Division Multiplexing

13.1 Introduction

Chapter 12 covered the basics of linear fiber-optic signal transportation. These are applicable to situations in which just a single optical signal is sent through each fiber. Fibers, however, are capable of carrying multiple independent optical signals on different wavelengths simultaneously with a minimum of mutual interaction. The general term for such shared use of fiber is *wavelength-division multiplexing* (WDM). A network design may choose to use WDM as an economical alternative to installing more fibers or as a means for combining signals that will be detected simultaneously by a common receiver. Both techniques have applications in cable systems.

This chapter will deal with the technology and performance issues encountered in various WDM applications, including component performance, mutual interaction in fibers, and link design trade-offs. You should first be familiar with the material covered in Chapter 12 because only the incremental information related to multiwavelength use of fibers will be covered here.

13.2 Wavelength Multiplexing: WWDM, CWDM, and DWDM

Wavelength multiplexing refers to any application in which multiple optical signals on different wavelengths share the use of common fibers. Within that general definition, however, there is a considerable range of applications and wavelength usage plans. Different acronyms are applied, somewhat inconsistently, to these plans to distinguish them. For our purposes, we will distinguish among three wavelength plans.

1310/1550 dual-wavelength plans — WWDM: The earliest WDM plan involved just two wavelengths: one in the 1310-nm window and one in the 1550-nm window. A typical cable application might involve transporting two signals over a shared link where they would be separated at the far end, or sending two signals modulated with nonoverlapping RF spectra to

a common detector where they would be detected and combined in a single operation. Although WDM generally refers to any level of multiplexing, the term is sometimes applied to 1310/1550 multiplexing, as distinguished from the more dense plans discussed later. ITU draft standard ITU-T G.671 considers any channel spacing greater than 50 nm to be *wide wavelength-division multiplexing* (WWDM); we will use that designation here.

20-nm-spaced plans — CWDM: An optical industry interim standard uses up to eight wavelengths, spaced 20 nm apart and centered approximately in the "third window" of optical fiber, also known as the "C band" (see Figure 12.6) at 1550 nm. The wavelengths are: 1470, 1490, . . . , 1610 nm. Generally, this scheme is referred to as *course wavelength-division multiplexing* (CWDM), in accordance with ITU-T G.671 (any channel spacing between 8 and 50 nm). ITU-T Recommendation G694–2, approved in June 2002, extends this down to 1270 nm (18 wavelengths), anticipating the ready commercial availability of fiber with no "water peak" of loss between the 1310-nm and 1550-nm transmission windows,[1,2] as discussed in Chapter 12. Such an extended-wavelength plan is, of course, applicable only to nonamplified systems, until such time as optical amplifiers with similarly extended bandwidths are developed.

Subnanometer-spaced plans — DWDM: The International Telecommunications Union (ITU) has defined a usage plan that can scale to as many as 45 wavelengths in the third window and whose spacings have been further split in some systems to yield twice that number. The defined channel designations are for channels spaced 100 GHz apart (about 0.8 nm). Regardless of whether 200-GHz or 100-GHz spacings are used, the usage plan is referred to as *dense wavelength-division multiplexing* (DWDM).

A few properties are common to all the plans, each with obvious parallels in RF technology.

- The closer the wavelengths are spaced, the harder (and more expensive) it is to separate them in the demultiplexers and simultaneously achieve adequate adjacent channel isolation, minimal in-channel flatness variation, and low insertion loss.
- The closer the wavelengths are spaced, the more frequency stability is required of the transmitters.
- The closer the wavelengths are spaced, the better the signal transmission velocities will match. Four-wave mixing is maximum when the signals are closely phase matched, whereas cross-phase modulation is maximum when group velocities are closely matched. The degree of matching is, of course, also dependent on fiber dispersion, with standard fiber having high dispersion at 1550 nm but low dispersion at 1310 nm. By contrast, close wave-

length spacing leads to reduced crosstalk from stimulated Raman scattering. These mechanisms are discussed later.

- The more wavelengths that share a fiber, the lower must be the power per wavelength for a given amount of mutual interaction due to nonlinear glass properties.

Cable systems using DWDM technology generally use 200-GHz-spaced channels from among the set of 20 listed in Table 13.1 For network designs that use fewer than 20 of these wavelengths, various vendors have chosen to offer different subsets of them. Most offer C21 through C35 as the first eight, but one vendor offers C39–C53 as the second eight, another offers C45–C59, and a third has chosen to offer C37–C51. This is obviously inconvenient to operators who wish to have multiple sources for optical transmitters and DWDM multiplexers.

Table 13.1 Commonly Used DWDM Channels

Wavelength (nm)	*ITU channel designation*
1560.61	C21
1558.98	C23
1557.36	C25
1555.75	C27
1554.13	C29
1552.52	C31
1550.92	C33
1549.32	C35
1547.72	C37
1546.12	C39
1544.72	C41
1542.94	C43
1541.35	C45
1539.77	C47
1538.19	C49
1536.61	C51
1535.04	C53
1533.47	C55
1531.90	C57
1530.33	C59

13.3 Components for WDM Systems

Constructing WDM systems requires the use of some components not required for single-wavelength links and tighter control over the specifications of other components.

13.3.1 Wavelength Multiplexers

Essential to shared use of fibers is a means by which to combine incoming signals at the transmit end and to separate them at the receiving end. It is possible to use simple wideband splitter/combiners or directional couplers to combine optical signals at the transmit end, and some applications do just that. The trade-off is that a broadband optical combiner, such as an RF combiner, has a minimal theoretical insertion loss of about 3 dB per two-way splitting level, with practical devices having losses 0.5–1.5 dB higher, whereas a two-wavelength multiplexer may have a loss of under 2 dB, and a 20-wavelength multiplexer a loss of under 4 dB.

At the current state of technology, the retail cost of a broadband combiner is about 10–15% of the cost of a 200-GHz-spaced WDWM multiplexer with an equivalent number of ports. Thus, the decision as to whether to use wavelength-specific or broadband combining at the transmit end of a WDM link must be driven by consideration of the overall link design, including the passive loss and the relative cost of obtaining higher-power transmitters as compared to the cost of wavelength-specific multiplexers.

At the receiving end, however, there is no alternative to using wavelength-specific demultiplexers if the signals are to be detected separately, because detectors are generally insensitive to minor changes in wavelength, with the result that all of the RF spectra modulated onto all of the received optical signals, along with beats among the optical spectra, will otherwise appear at the detector output port.

The technology used to build WDM filters is developing rapidly, with corresponding improvements in both performance and price.[3] Typical 16- or 20-wavelength, 200-GHz channel-spacing filters available in 2002 had the following characteristics in devices costing about $500 per wavelength:

Adjacent channel isolation	27–30 dB
Nonadjacent channel isolation	30–40 dB
Total insertion loss, including connectors	3.8–4.0 dB
Differential insertion loss	1.0 dB
In-channel loss variation	0.5 dB
Polarization-dependent loss variation	0.5 dB

Not generally specified is the maximum in-channel transmission slope, which is crucial to calculating how various phase modulation mechanisms convert to noise and second-order distortion.

Both the price and performance parameters have significantly improved recently, and they are expected to continue to improve.

13.3.2 Gain-Flattened Optical Amplifiers

Not only fibers, but optical amplifiers, are capable of handling multiple optical signals on separate wavelengths. In single-wavelength applications, they are almost always operated in a saturated mode and thus have a constant output power regardless of drive level (over a defined range, of course). Thus, any variation of gain with wavelength is not an important consideration. When operated in saturated mode while amplifying multiple wavelengths, however, the total output power will be divided among all the signals, so if one input level changes or a signal is removed or added, the level of all the other signals will also change, which is obviously unacceptable in analog optical links where the RF output power of a detector is a function of both modulation level and optical power level.

Manufacturers avoid this problem by offering products that can optionally be operated in constant-gain-per-wavelength mode. As with coaxial amplifiers, however, the uniformity of gain with (optical) frequency is an essential factor. In general, amplifiers designed to handle a single wavelength do not exhibit a sufficiently flat response. Not only that, but their optical frequency response varies as a function of input level. Thus, a class of amplifiers known as *gain-flattened* amplifiers has been developed for DWDM applications. These devices are designed to operate in a fixed-gain mode of operation and offer a much flatter frequency response. Typical specifications for such a device include gain flatness within 1 dB peak-to-peak from 1530 to 1565 nm and over a composite power input range of −6 dBm to +6 dBm. Typical devices offer a maximum composite power output of up to 20 dBm per-wavelength gain of 17–26 dB, and a noise figure of 5–5.5 dB.[4]

The imperfect frequency response of the optical channel (the total of multiplexer, demultiplexer, and amplifier variation) interacts with each source's incidental wavelength modulation (chirp) to produce distortion products. To the extent that the response variation is approximately linear across the transmitter modulated linewidth (the usual case), the result will be a degradation in CSO in the demodulated signal. If the response variation is noticeably nonlinear, third-order products will be produced as well.

Second, the imperfect in-channel frequency response will interact with any cross-phase modulation occurring before the device to produce cross-amplitude modulation, as discussed later.

Finally, the broader response variation (the total of wideband variation in amplifier response, multiplexer and demodulator channel-to-channel insertion loss variations) will also make the various optical signals arrive at their detectors

at different levels, resulting in a less-than-optimum balance between noise and distortion for some of the wavelengths.

13.4 WDM-Specific Design Factors

Figure 13.1 illustrates the two generic applications of WDM. In the first application, optical signals on multiple wavelengths are generated by independent transmitters, multiplexed together, transmitted through a shared fiber, optionally amplified, demultiplexed at the receiving end, and detected by independent receivers. In the second application, the signals are not demultiplexed, but rather fed to a common receiver.

A typical example of the first application would be the transmission of multiple node-specific digital programming from a headend to a hub, where the signals destined for each node would be separated before detection. The second application is frequently used for combining common RF spectra modulated on one wavelength (for instance, 50–550 MHz modulating a 1310-nm transmitter) with node-specific programming (using some portion of the spectrum above 550 MHz modulating a 1550-nm transmitter) and then detecting both at the node using a common detector. Sometimes these applications are cascaded, as will be discussed in Chapter 18.

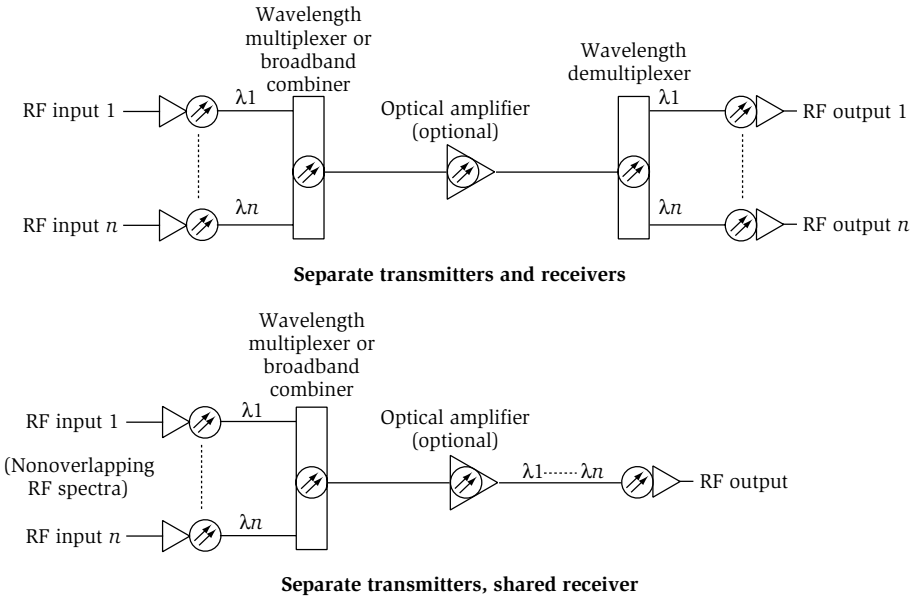


Figure 13.1 Separate versus shared receiver WDM applications.

In the first application, the following performance parameters must be considered, in addition to those discussed in Chapter 12:

- Detectors are largely wavelength independent. Thus, to the extent that light from other wavelengths is not perfectly excluded from the desired wavelength output of the demultiplexer, the recovered RF spectrum will also contain some level of the signals modulated on other wavelengths.
- As discussed in Chapter 12, optical signals parametrically modulate the properties of the glass through which they pass. More specifically, the refractive index varies slightly in proportion to the instantaneous optical power level (known as the *optical Kerr effect*, or OKE). In multiple-wavelength systems, these variations lead to various interwavelength effects: *cross-phase modulation* (XPM), *four-wave mixing* (FWM), and *cross-polarization modulation*. These, in turn, interact with the properties of both the fiber and the discrete devices along the transmission path to produce noise and cross modulation in the detected RF signals. Additionally, the presence of each signal can cause the fiber to have incremental gain or loss through a process known as *stimulated Raman scattering* (SRS).
- To the extent that the optical frequency response of each channel (including the multiplexer, any amplifiers, and the demultiplexer) is not flat, that will combine with any transmitter chirp to generate in-band second-order distortion products in the recovered RF spectrum.

When a common detector is fed more than a single wavelength, some additional parameters must be considered.

- Since there is no way to filter the RF output spectra before combining, each portion of the final spectrum will be degraded by broadband noise from both optical links.
- Similarly, each portion of the spectrum will be degraded by distortion components arising from the other links that fall in the RF spectrum of the first link.
- The final balancing of RF levels between those transmitted on each link will be a combination of the *optical modulation index* (OMI) and relative optical received levels and thus cannot be adjusted at the receiving site.

Each of these will be considered in detail later.

13.5 Crosstalk Mechanisms[5]

The first set of factors to be considered are those that lead to crosstalk — defined as the level of postdetection products and wideband noise relative to desired signals that are caused by the presence of modulated optical signals other than the one

whose performance is being considered and that share use of a common network segment that may include multiplexers, fiber, amplifiers, and/or demultiplexers.

13.5.1 Imperfect Demultiplexer Wavelength Isolation

Imperfect multiplexer isolation affects the amount of light from each source that shows up at the output of other sources, but otherwise it has no effect on link operation. Imperfect demultiplexer isolation, however, makes some light from undesired wavelengths impinge on the detector along with the light at the desired wavelength.

In the detector, each of the incoming optical signals will be detected, resulting in a composite RF spectrum that contains elements of the modulating spectra of all the signals. Since the detector is a square-law device, the contributions from undesired signals will be lower than the desired RF output by twice the difference in optical levels at the detector input, assuming similar received optical power levels and modulation indices (the assumption throughout this chapter, unless stated otherwise).

Thus, if the adjacent channel isolation in a DWDM demux is 30 dB and OMIs are similar, then the modulating signal from each of the adjacent wavelengths will appear about 60 dB below the desired signal after detection. If the modulation is analog video and the same nominal channel frequencies are used for each link, then the undesired video carriers will appear close to the desired carriers and have the same effect as ingressing signals. If the modulation on both desired and adjacent signals is digital, then the undesired RF output will appear noiselike and the two adjacent signals together will generate a contribution to the total link C/N of 57 dB.

More generally, if the link were carrying a total of n wavelengths, each modulated with a similar spectrum of signals, the adjacent channel isolation were A dB, and the nonadjacent isolation were B dB, then the total link C/I contribution (for all but the shortest and longest wavelengths) due to imperfect demultiplexer isolation would be

$$C/I_{\text{ISOLATION}} = -10\log[2 \times 10^{-A/5} + (n-3) \times 10^{-B/5}] \tag{13.1}$$

For the highest and lowest wavelengths, there is only one adjacent wavelength, so the equation becomes

$$C/I_{\text{ISOLATION}} = -10\log[10^{-A/5} + (n-2) \times 10^{-B/5}] \tag{13.2}$$

where

$C/I_{\text{ISOLATION}}$ = the ratio of the desired to the undesired RF signal powers in the demodulated spectrum of the victim optical carrier, in dB

A = the adjacent optical channel isolation of the WDM demultiplexer in dB

$B =$ the nonadjacent optical channel isolation of the WDM demultiplexer in dB

$n =$ the number of optical carriers

Assuming an environment where all optical signals are at the same nominal level and optical modulation index (the usual case), the level of postdetection crosstalk interference will be independent of the average optical levels and modulation frequency; thus, this is considered a linear degradation factor.

13.5.2 Cross-Phase Modulation Combined with Fiber Dispersion (XPM-D)

Recall from Chapter 12 that self-phase modulation (SPM) results from the fact that the refractive index of optical fiber changes slightly in the presence of high electrical fields. Thus, a sufficiently high amplitude signal will cause the index of refraction to vary as the square of its own instantaneous electrical field strength, resulting in incidental phase modulation, because the velocity of propagation varies inversely with the index of refraction.

When two or more optical signals share a fiber, whatever modulation of the index of refraction takes place affects all of the signals. Thus, the amplitude variations of each of the individual signals result in some degree of phase modulation of the other signals. In a perfect transmission path and with a perfect broadband detector, this phase modulation would have no effect on performance. The cross-phase modulated signal, however, travels through fiber that exhibits chromatic dispersion to create cross-intensity modulation in the optical signal and, thus, cross-amplitude modulation in the detected RF signals.

Assuming equal optical modulation indices, crosstalk from each interfering modulated optical carrier due to the interaction of cross-phase modulation with fiber dispersion is given by

$$C/I_{XPM-D}(\text{dB}) = -20\log\left(\frac{4\pi n_2 \beta \Omega^2 P_{20} \rho_{XPM}}{\lambda_2 A[\alpha^2 + (d_{12}\Omega)^2]}\right) - 10\log(1 + e^{-2\alpha z} - 2e^{-\alpha z}(1 - \alpha z) \cos(d_{12}\Omega z) - 2z[\alpha + d_{12}\Omega e^{-\alpha z}\sin(d_{12}\Omega z)] + [\alpha^2 + (d_{12}\Omega)^2]z^2) \tag{13.3}$$

where all variables are in a consistent set of units. In meter-kilogram-second (MKS) units:

$C/I_{XPM} =$ the ratio, in dB, of desired to undesired RF signal powers in the demodulated spectrum of the victim optical carrier, in dB (assuming equal optical modulation indices of both optical signals)

$n_2 =$ the nonlinear refractive index of the fiber, typically $2.6 \times 10^{-20}\ \text{m}^2/\text{W}$

$\beta = -(\lambda_2{}^2 D)/(2\pi c)$

λ_1 = the wavelength of the victim optical carrier, generally between 1.530×10^{-6} and 1.560×10^{-6}m

λ_2 = the wavelength of the interfering modulated optical carrier, in meters

D = the dispersion coefficient of the fiber in sec/m^2. For standard fiber near 1550 nm, this is typically 17 ps/nm-km = 17×10^{-6}sec/m^2

c = the speed of light in a vacuum = 3×10^8 m/sec

Ω = the frequency of the modulating RF signal in radians = $2\pi f_I$, where f_I is the modulating frequency in Hz

P_{20} = the power level of the modulated interfering optical signal, in watts

ϱ_{XPM} = the effective polarization overlap between the interfering and the victim optical carriers. This varies from 1 for copolarized signals to 1/3 for cross-polarized signals. At 45° it is 2/3

A = the effective mode cross-sectional area of the fiber core, about 80×10^{-12} m^2 for non-dispersion-shifted fibers and 50×10^{-12} m^2 for dispersion-shifted fibers

α = the fractional power attenuation of the fiber per meter of length

= $1 - 10^{-(\alpha 0/10{,}000)}$, where $\alpha 0$ is the attenuation in dB/km. A typical value of $\alpha 0$ for standard fiber at 1550 nm is 0.21 dB/km, for which $\alpha = 4.835 \times 10^{-5}$per meter

z = the length of the shared fiber, in meters

d_{12} = the group velocity mismatch between the interfering and victim optical carriers $\approx D(\lambda_1 - \lambda_2)$

Converting this to more common engineering units and assuming operation near 1550 nm, we get

$$C/I_{XPM-D}\ (\text{dB}) = 459.5 - 2P_I - 20\log\left(\frac{f_{RF}^2 DL^2 \rho_{XPM}}{A[X^2 + Y^2]}\right) - 10\log(1 + e^{-2X} - 2e^{-X} (1 - X)\cos(Y) - 2X - 2Ye^{-X}\sin(Y)] + X^2 + Y^2) \quad (13.4)$$

where

P_I = the launch power of the interfering optical carrier in dBm

f_{RF} = the modulation frequency in MHz

D = the fiber dispersion in ps/nm-km. For non-dispersion-shifted fiber near 1550 nm this is typically 17

L = the length of the fiber in km

ϱ_{XPM} = the effective polarization overlap between the interfering and victim optical carriers. This varies from 1 for copolarized signals to 1/3 for cross-polarized signals and is 2/3 for 45° relative polarization

A = the effective cross-sectional area of the fiber core. For non-dispersion-shifted fiber this is typically 80×10^{-12} m^2

α = the fractional power attenuation of the fiber per km

= $1–10^{-(\alpha 0/10)}$, where $\alpha 0$ is the attenuation in dB/km. A typical value of $\alpha 0$ for standard fiber at 1550 nm is 0.21 dB/km, for which $\alpha = 0.047$

$\Delta\lambda$ = the difference in wavelength between the interfering and victim optical carriers in nm

X = the "loss parameter" αL. For fiber with a loss of 0.21 dB/km, this is $0.047L$, where L is the length of the fiber in km

Y = the "walkoff parameter" $D\,\Delta\lambda(2\pi f_{RF})L$. For fiber with a dispersion of 17 ps/nm-km, this is $1.068 \times 10^{-4}\Delta\lambda\, f_{RF}L$.

The amount of optical cross-phase modulation, and thus intensity modulation after interaction with the dispersion, varies linearly with the level of the interfering optical signal. Thus, after detection, this degradation factor will vary 2 dB for every 1-dB change in the level of the interfering optical signal.

The amount of optical cross-phase modulation also varies according to the polarization match between the interfering and victim optical signals in the fiber. For cross-polarized signals, the cross modulation is one-third that for polarization-aligned signals.

Due to dispersion in the fiber, the interfering and victim optical signals will travel at slightly different velocities. This property, also known as *walk-off*, reduces the peak cross-phase modulation because the peak amplitude of the interfering signal "slides" along the victim signal rather than synchronously acting on the same spot in time. The greater the wavelength difference, the faster the walk-off occurs, resulting in less peak cross-phase modulation.

The effects of cross-phase modulation are greatest at higher RF frequencies. One reason is that a constant level of phase deviation versus frequency results in an optical frequency deviation that linearly increases with modulating frequency, and it is the frequency change that interacts with dispersion to create the crosstalk.

Figure 13.2 shows the maximum and minimum expected crosstalk from this mechanism for a single copolarized interfering carrier whose launch power is +7 dBm and whose optical carrier is 1.5 nm away from the victim carrier, with both transmitted through 30 km of standard fiber.

13.5.3 Stimulated Raman Scattering (SRS)

At a sufficiently high optical signal level, one optical signal can act as a "pump" so as to provide gain (either positive or negative) to other signals sharing the fiber. If the

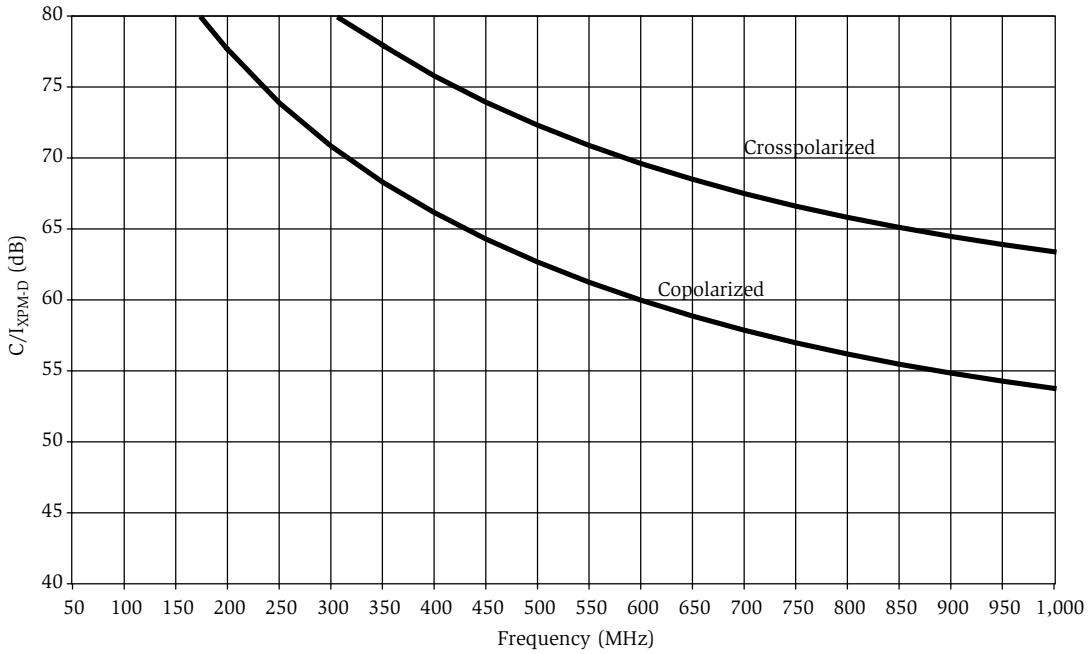


Figure 13.2 C/I due to cross-phase modulation interacting with fiber dispersion.

pump signal has a wavelength that is shorter than the victim signal, then the gain will be positive, whereas if the pump signal is longer, the gain will be negative (in other words, loss). The gain is maximum for WDM channels with optical frequencies separated by about 13 THz and is proportionately lower for narrower separations. Another way of looking at this phenomenon is that power is transferred from the shorter-wavelength optical carrier to the longer-wavelength carrier as they travel through the shared fiber; but since they are both modulated, the SRS gain follows the instantaneous carrier levels and results in mutual cross-amplitude modulation.

Assuming equal optical modulation indices, crosstalk from each interfering modulated optical carrier due to stimulated Raman scattering is given by

$$C/I_{SRS}\,(\text{dB}) = -10\log\left[\left(\frac{\rho_{SRS} g_{12} P_{20}}{A}\right)^2 \left(\frac{1 + e^{-2\alpha z} - 2e^{-\alpha z}\cos(\Omega d_{12} z)}{\alpha^2 + (\Omega d_{12})^2}\right)\right] \quad (13.5)$$

where the units are the same as for Equation (13.3), with the following additions:

C/I_{SRS} = the ratio of desired to undesired RF signal powers in the demodulated spectrum of the victim optical carrier, in dB (assuming equal optical modulation indices of both optical signals)

ρ_{SRS} = the effective polarization overlap between the interfering and victim signals. This equals 1 for copolarized signals and less than 0.1 for cross-polarized signals

g_{12} = the Raman gain coefficient, which is positive if the interfering signal is of shorter wavelength than the victim signal and negative if the interfering signals is of longer wavelength $= g_{SRS}c(\lambda_2 - \lambda_1)/(\lambda_2\lambda_1)$, where g_{SRS} = the Raman gain slope for the fiber, typically 5×10^{-27} m/W-Hz.

Converting this to more convenient engineering units, we get

$$C/I_{SRS}(\text{dB}) = 304.1 - 2P_I - 20\log\left(\frac{\rho_{SRS}\Delta\lambda L}{A}\right) - 10\log\left(\frac{1 + e^{-2X} - 2e^{-X}\cos(Y)}{X^2 + Y^2}\right) \tag{13.6}$$

where the units are the same as for Equation (13.4), with the following additions:

C/I_{SRS} = the ratio of desired to undesired RF signal powers in the demodulated spectrum of the victim optical carrier, in dB (assuming equal optical modulation indices of both optical signals)

ϱ_{SRS} = the effective polarization overlap between the interfering and the victim optical carriers. This varies from 1 for copolarized signals to less than 1/10 for cross-polarized signals. At 45° relative polarization, it has a value of 0.5.

As with XPM crosstalk, SRS varies 2 dB for every 1-dB increase in optical carrier level; however, the degree of SRS "coupling" between the signals is more highly dependent on the polarization match between them and drops to less than a tenth for cross-polarized signals, meaning that the RF crosstalk can vary more than 20 dB.

Crosstalk due to SRS is maximum at low modulation frequencies but decreases in a nonlinear way, with a periodic variation superimposed on the general decrease as a function of frequency due to the walkoff between the interfering and victim channels. The variation exhibits more cycles across the modulation frequency spectrum when the optical channels are more widely separated, since the velocity mismatch is then greater.

Figure 13.3 shows the maximum and minimum expected crosstalk from SRS for a single copolarized interfering carrier whose launch power is +7 dBm and whose optical carrier is 24 nm away from the victim carrier, with both transmitted through 30 km of standard fiber.

SRS crosstalk is typically the most serious of the crosstalk mechanisms at far wavelength spacings and low modulation frequencies, with the crosstalk at 50 MHz being worse than for any other mechanism at any other frequency. One researcher has suggested reducing the effect of SRS by transmitting pairs of closely spaced optical carriers, modulated 180° out of phase, so that the SRS

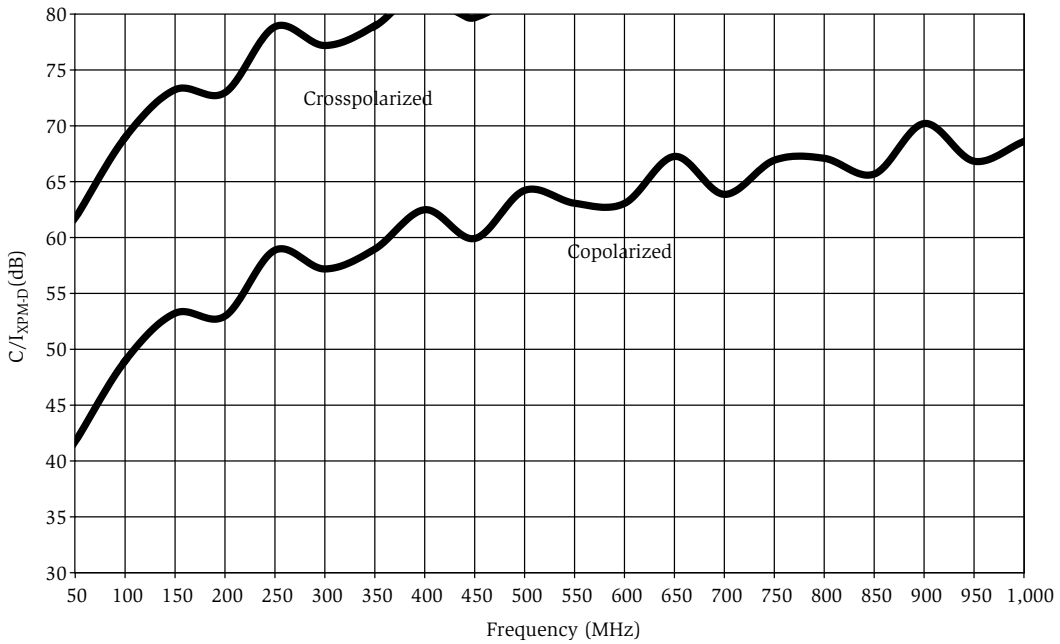


Figure 13.3 C/I due to stimulated Raman scattering.

crosstalk on a third carrier from one would approximately cancel the crosstalk from the other. Experimental data confirms a reduction of about 30 dB at 50 MHz and 15 dB at 800 MHz for 1-nm-spaced interfering carriers, with 4- to 5-nm spacing to the victim carrier. Whether this will be a cost-effective solution to the problem has yet to be determined.[6]

13.5.4 Cross-Phase Modulation Combined with Transmission Slope (XPM-TS)

Aside from the conversion from XPM to intensity modulation that takes place in the fiber itself, the phase-modulated victim signal also reacts with any device whose transmission slope is not flat with frequency (*transmission slope*, or TS). Generally, the least flat component is the wavelength demultiplexer. This gives rise to a slope-detection conversion from phase to amplitude modulation. Assuming equal modulation of the two carriers, the crosstalk is given by this equation:

$$C/I_{XPM-TS}(\text{dB}) = -20\log\left(\left[\frac{1}{T_0}\frac{\partial T}{\partial f}\right]\frac{2n_2\Omega P_{20}\rho_{XPM}}{\lambda A}\right) - 10\log\left(\frac{1+e^{-2\alpha z}-2e^{-\alpha z}\cos(d_{12}\Omega z)}{[\alpha^2+(d_{12}\Omega)^2]}\right) \tag{13.7}$$

where all variables have the same units and meanings as in Equation (13.3), with the following additions:

T_0 = the linear transmission coefficient of the component in question (i.e., the ratio of output power to input power)

$\partial T/\partial f$ = the differential of the linear transmission coefficient at the wavelength being used with respect to frequency (i.e., the change in the ratio of output power to input power divided by the change in optical frequency)

The product of these, in MKS units, can be calculated using

$$\frac{1}{T_0}\frac{\partial T}{\partial f} = 1 - 10^{-S/10^{-9}} \tag{13.8}$$

where S is the transmission slope of the demultiplexer or other component in dB/GHz.

In more common engineering units, this reduces to

$$\begin{aligned} C/I_{XPM-TS}\,(\mathrm{dB}) = 313.5 - 2P_I - 20\log\left(\frac{S_T f_{RF}\rho_{XPM}L}{A}\right) \\ - 10\log\left(\frac{1 + e^{-2X} - 2e^{-X}\cos(Y)}{X^2 + Y^2}\right) \end{aligned} \tag{13.9}$$

where the units are the same as for Equation (13.4), with the following additions:

$$S_T = 1 - 10^{-S/10}$$

L = the length of the fiber in km

Unfortunately, the in-band transmission slope of wavelength demultiplexers is seldom specified. One researcher has found typical devices to have a slope that averaged 0.02 dB/GHz but were as high as 0.11 dB/GHz.[7]

As with crosstalk, due to the interaction between cross-phase modulation and fiber dispersion, XPM-TS is greater at higher modulating frequencies and higher optical powers but is reduced at wider optical channel spacing and higher fiber dispersion. Figure 13.4 shows the variation with frequency for two copolarized signals spaced by 1.5 nm and launched at 7 dBm. Curves are plotted for 0.02 dB/GHz and 0.1 dB/GHz, which represent the typical range of measured values for commercially available DWDM demultiplexers.

XPM-TS can add either in phase or out of phase with XPM-D, depending on the slope of the filter loss.

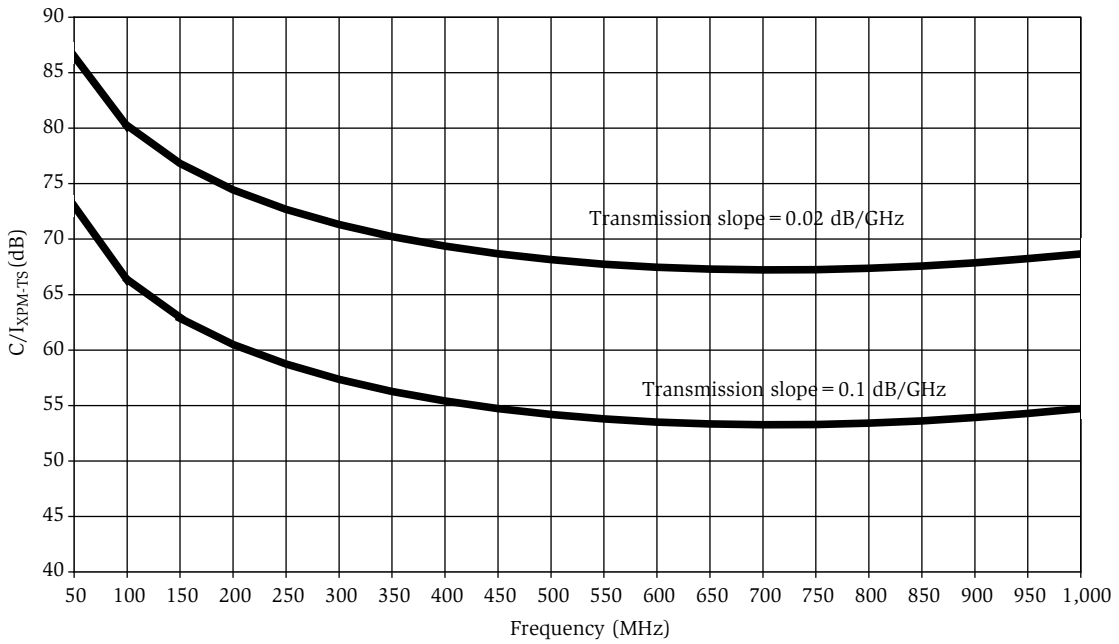


Figure 13.4 C/I due to cross-phase modulation interacting with imperfect demux channel flatness.

13.5.5 Optical Kerr Effect Combined with Polarization-Dependent Loss (OKE-PDL)

As seen already, both XPM and SRS are maximum when the polarizations of the interfering signals are aligned with the target signal and minimum when they are cross polarized. When they are at some other relative polarization, however, the polarization, as well as the gain and phase of the target signal are modulated.

This can be understood if the target signal is viewed as the vector addition of two signals, one polarization aligned with the interfering signal and one cross polarized to it. The one that is polarization aligned undergoes much greater cross-amplitude and cross-phase modulation than the one that is cross polarized, with the result that the vector sum of the two has a net polarization that varies.

As with XPM, in a perfect network followed by a perfect detector, the OKE would have no effect on the detected signals. When either the fiber or a discrete device has a *polarization-dependent loss* (PDL), however, it is converted to amplitude modulation. One researcher found a typical WDM demultiplexer to have a PDL of 0.03–0.17 dB[8]; however, as noted earlier, typical specifications for production devices are of the order of 0.5 dB.

Assuming equal optical modulation indices, crosstalk from each interfering modulated optical carrier due to the interaction of the optical Kerr effect with polarization-dependent loss in a terminating component is given by the following general equation:

$$C/I_{OKE-PDL}\,(\mathrm{dB}) = -10\log\left[\left(\frac{\Delta T \rho_3 t_3}{3(1-\frac{\Delta T}{2}(1-t_2))}\right)^2\left(\frac{4\Pi n_2 P_{20}}{\lambda A}\right)^2 \left(\frac{1+e^{-2\alpha z}-2e^{-\alpha z}\cos(d_{12}2\pi f_{RF}L)}{\alpha^2+(d_{12}2\pi f_{RF})^2}\right)\right] \tag{13.10}$$

where the units are the same as for Equation (13.3), with the following additions:

$C/I_{OKE-PDL}$ = the ratio of desired to undesired RF signal powers in the demodulated spectrum of the victim optical carrier, in dB, for a single interfering carrier

ΔT = the transmission difference through the terminating device as a percentage of the maximum transmission as a function of polarity

$= 1-10^{-\mathrm{PDL}/10}$, where PDL is the polarity-dependent loss through the device in dB

ϱ_3 = the effective polarization overlap factor for this mechanism. This varies from 0.75 at 45° relative polarization to zero for signals that are either copolarized or orthogonally polarized

t_2, t_3 = transmission parameters that can vary over the range from +1 to −1. Crosstalk is maximum for $t_3 = 1$ and $t_2 = -1$.

Converting this to more common engineering units and calculating the maximum possible crosstalk, we get

$$C/I_{OKE-PDL}(\mathrm{dB}) = 263 - 2P_I - 20\log\left[\frac{\Delta T L \rho_3}{A}\right] - 10\log\left[\frac{1+e^{-2X}-2e^{-X}\cos(Y)}{X^2+Y^2}\right] \tag{13.11}$$

where the units are the same as for Equation (13.4), except that ΔT and ρ_3 are as in Equation (13.9).

Although the maximum possible OKE-PDL effect can be calculated, the level in practical applications can vary widely, due to two factors. First, the degree of cross-polarization modulation varies with the relative polarizations between the interfering and victim signals. Second, the conversion from polarization modulation to amplitude modulation depends on the polarization of the victim signal

relative to the maximum slope of loss versus polarization in the demultiplexer. One researcher found that the experimental C/I from this effect was 6 dB better than the minimum predicted and that the median C/I was 9 dB better.[9]

With good-quality WDM demultiplexers, the total crosstalk between optical channels is dominated by SRS effects at low modulation frequencies and by XPM effects at high modulation frequencies. In the region where those effects would be expected to cancel through vector addition, however, the OKE-PDL effect becomes dominant. Generally this occurs when optical channels are closely spaced, when the interfering optical carrier is at a shorter wavelength than the victim carrier (so that the phase of the SRS effect is opposite that of the XPM effect), and at the low end of the RF spectrum. Above 0.1 dB of PDL, OKE-PDL can be the dominant crosstalk mechanism at 50 MHz, above 0.4 dB it can dominate across most of the typical CATV spectrum if the polarizations between interfering and target signals happen to be at or close to 45° polarization difference and the output is oriented to the maximum PDL slope of the terminating device. Unlike the other crosstalk mechanisms, OKE-PDL effects typically vary with time and temperature because the polarization in field-installed single mode fiber is essentially uncontrolled and, thus, the polarization of the received signal will vary with respect to the slope of the PMD of the terminating filter. In field testing, OKE-PDL varied over as much as 30 dB, approximating the theoretical maximum at times.[10] Finally, as with XPM-TS, OKE-PDL will tend to add either constructively or destructively with the XPM-D and SRS, depending on the slope of the PDL effect, which is not predictable in practical situations.

Figure 13.5 plots the maximum possible OKE-PDL for two optical signals spaced 10 nm apart, polarized 45° relative to each other, and launched through 30 km of fiber at 7 dBm.

13.5.6 Four-Wave Mixing

The fact that the refractive index responds to the total optical power level means that the transmission system is nonlinear. As with nonlinearity in an RF amplifier, this leads to the creation of new optical signals whose frequencies are sums and differences of the primary signal frequencies. The third-order products (equivalent to CTB products at RF) can fall in the wavelength range of interest and, given the equal spacing of channels in both CWDM and DWDM systems, can fall nominally on top of desired signals. These are at frequencies $A + B - C$ and $2A - B$, where A, B, and C are optical frequencies of the primary signals. As expected, the buildup of products is exactly the same as for CTB products in RF systems. In optical systems, these are known as *four-wave mixing* (FWM) products, because three wavelengths (A, B, and C) combine to produce a fourth wavelength.

Four-wave mixing products can have two types of effects. If the products fall within the bandwidth of the wavelength demultiplexer channel, then they will pass through the same demultiplexer port and be demodulated along with the desired signals, with the resultant RF products being suppressed below the

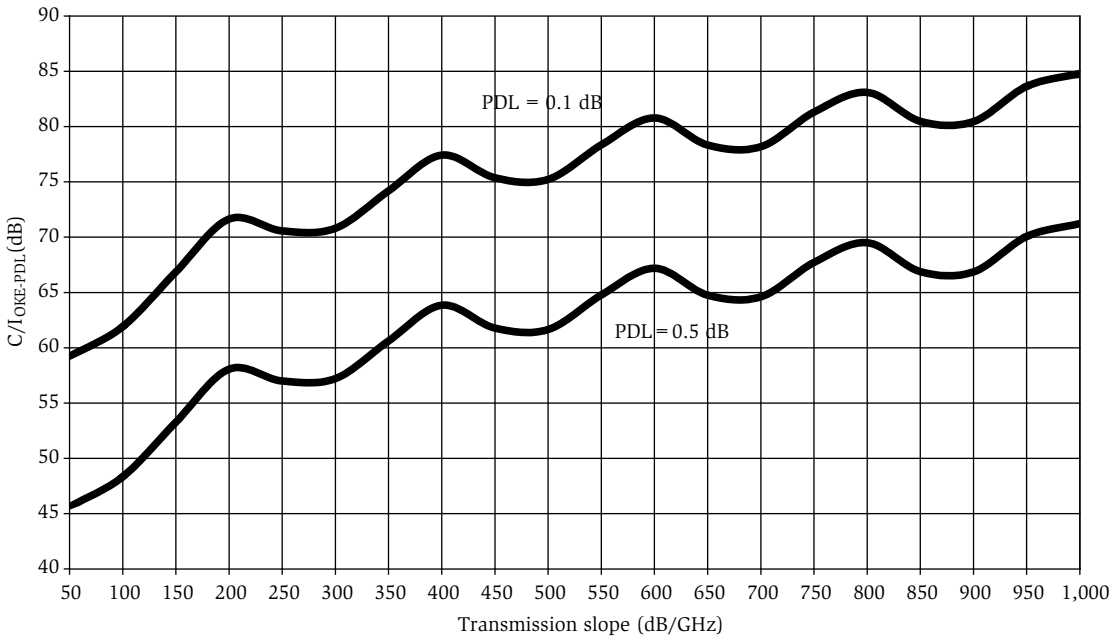


Figure 13.5 C/I due to OKE or cross-polarization modulation interacting with polarization-dependent loss.

desired RF spectra by twice the difference in optical levels. Should the difference in the optical frequencies between the desired signal and a third-order product be less than the highest modulating RF frequency, however, then a high-level beat product will be generated in the demodulated RF spectrum. This is obviously of greater concern when the transmitters are externally modulated units whose optical frequencies are very stable. It is less of a concern with directly modulated DFB transmitters, due to chirp.

Four-wave mixing is greatest when the signals giving rise to a product interact cohesively over an extended length of fiber. When the signals travel at different velocities, the interaction is reduced. Therefore, it is of greatest concern when dispersion-shifted fiber is used and when the optical channels are closely spaced. For baseband digital systems utilizing channels spaced as closely as 50 GHz and dispersion-shifted fiber, it can be the dominant distortion mechanism. The efficiency of the process drops dramatically, however, when the interfering signal moves by as little as 100–200 GHz from the zero-dispersion wavelength, falling off as the square of the dispersion and as the fourth power of the difference in wavelengths of the interfering and victim signals.[11,12] This effect is negligible for the most common cable television use of DWDM, which utilizes 200-GHz-spaced channels transmitted through standard optical fiber.

13.5.7 Summary of Cross-Modulation Effects in WDM systems

All of the cross-modulation mechanisms described here are coherent for a given pair of wavelengths, and thus the effects add vectorially. When the interfering signal is at a longer wavelength than the victim signal, XPM-D and SRS effects are at nominally opposite phases and the net crosstalk is less than the larger of the two. When the interfering signal is at a shorter wavelength, they add to produce crosstalk greater than either mechanism alone. On the other hand, SRS crosstalk increases, while XPM decreases, as the difference in wavelengths increases, resulting generally in a broad minimum as the wavelength spacing approaches zero, whereas OKE-PDL has a broad maximum at small wavelength spacings and prevents XPM and SRS effects from cancelling completely at any spacing.

XPM-TS, like OKE-PDL, can add constructively or destructively with the net of XPM and SRS, depending on the slope of the response of the device, which is generally not predictable. Finally, crosstalk due to imperfect demultiplexer isolation will add constructively with XPM-D crosstalk.

Thus, the most conservative calculation of optical cross modulation in a two-wavelength system would require adding the worst-case demultiplexer isolation and XPM-D magnitudes, then adding or subtracting (depending on relative wavelengths) the SRS magnitude, and then adding the XPM-TS and OKE-PDL magnitudes to the absolute value of the previous calculation. More formally:

$$C/I_{\text{CROSSTALK}}\,(\text{dB}) = -20\log\Big[\big|10^{-I/20} + 10^{-C/I_{XPM-D}/20} \pm 10^{-C/I_{SRS}/20}\big| + 10^{-C/I_{OKE-PDL}/20} + 10^{-C/I_{XPM-TS}/20}\Big] \tag{13.12}$$

where the plus sign is used where the interfering carrier is at a shorter wavelength than the victim carrier and the minus sign otherwise and

$C/I_{\text{CROSSTALK}}$ = the net crosstalk occurring between two optical carriers similarly modulated

I = the isolation of the WDM demultiplexer at the wavelength of the interfering carrier, in dB

C/I_{XPM-D} = the crosstalk due to cross-phase modulation interacting with fiber dispersion, in dB

C/I_{SRS} = the crosstalk due to stimulated Raman scattering, in dB

$C/I_{OKE-PDL}$ = the crosstalk due to the optical-Kerr-effect-caused cross-polarization modulation interacting with polarization-dependent loss in the WDM demultiplexer, in dB

C/I_{XPM-TS} = the crosstalk due to cross-phase modulation interacting with imperfect channel flatness in the WDM demultiplexer, in dB

Such a calculation will overstate the total cross modulation if worst-case levels are assumed for all mechanisms, because the OKE-PDL magnitude cannot be maximum at the same input-signal relative polarization that maximizes SRS and XPM. To be safe, the calculation can be run with relative polarizations of 0° (where OKE-PDL is zero) and also at 45° (where OKE-PDL is maximum) to see which results in the worst-case crosstalk.

In practical headend and field installations, relative polarizations will not be known or controlled and can be assumed to vary randomly. Since XPM crosstalk varies by about 10 dB with polarization and SRS varies by 20 dB or more, at best, an approximation of the total likely cross modulation will be possible.

In systems of more than two wavelengths, the calculated cross modulations between various pairs of optical carriers will vary due to different channel spacings, different numbers of lower and higher interfering channels relative to each victim channel, different qualities of optical demultiplexer channels (combined with transmitter wavelength tolerance relative to nominal channel center wavelength), and random polarization combinations. In addition, it is not uncommon in CATV systems for at least part of the spectrum to be common to multiple transmitters driving a common fiber. This means that, potentially, multiple interfering carriers could be modulated synchronously over at least part of the RF spectrum and therefore that the cross-modulation effects could add vectorially rather than randomly. One researcher has found that ignoring possible synchronous effects (that is, by adding the calculated distortion from each interfering carrier on a power, rather than voltage, basis) agreed with experimental results within 7 dB at all frequencies and within 3 dB above 500-MHz modulation in an eight-wavelength system.[13]

Using this approach, Figure 13.6 shows the maximum possible contributions of each crosstalk mechanism and the net crosstalk affecting the longest and shortest wavelengths for a 16-wavelength, 30-km-long optical circuit, with each optical carrier launched at a power of +7 dBm. The optical carriers were assumed to occupy odd-numbered ITU channels C21 through C51 and to be copolarized. Fiber was assumed to be standard unshifted fiber with a dispersion of 17 ps/nm-km and a loss of 0.21 dB/km. The wavelength demultiplexer was assumed to have a transmission slope of 0.05 dB/GHz, a PDL of 0.25 dB, an adjacent channel isolation of 27 dB, and a nonadjacent channel isolation of 32 dB. The optical carriers were assumed to be modulated to similar levels, with crosstalk evaluated over a range extending from 50 to 1000 MHz.

The composite crosstalk is seen to be dominated by factors related to the quality of the demultiplexer, except for frequencies below 500 MHz, where SRS effects are largest. As would be expected, the shortest-wavelength signal (C51) has better performance, due to the out-of-phase combining of SRS with the sum of isolation and XPM-D. Since the signals are copolarized, OKE-PDL does not contribute to the composite crosstalk level.

Figure 13.7 is the same, except evaluated for one of the two center optical channels (C35). The performance is better at low modulation frequencies

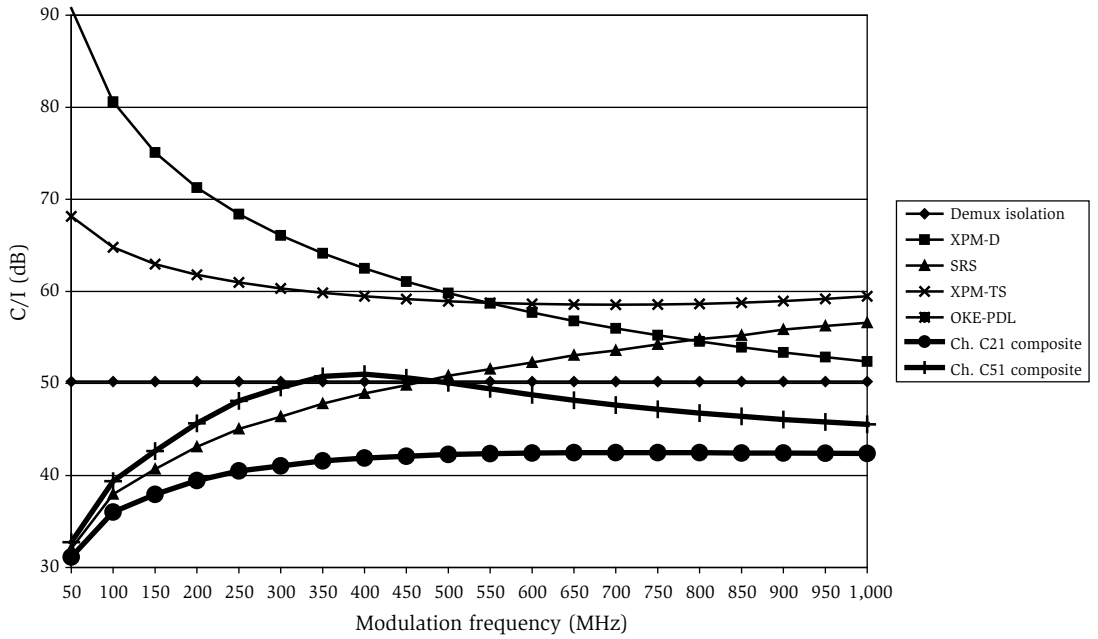


Figure 13.6 Maximum composite C/I: 16 wavelengths Copolarized, shortest and longest wavelengths.

because there are fewer widely spaced channels and therefore less SRS. At the high end of the spectrum, the cross modulation is slightly worse because of a larger contribution from the demultiplexer isolation due to the presence of two adjacent channels, rather than one as in the previous graph.

Figure 13.8 is the same as the first graph, but it shows the change if we assume that the signals are polarized at 45°, where OKE-PDL is maximized, but that XPM and SRS are both reduced. In this case, the crosstalk is dominated by OKE-PDL and demux isolation, and the composite C/I is degraded by 2–3 dB across most of the spectrum.

In summary, in every case modeled, the performance is dominated over at least the top half of the spectrum by some performance limitation of the demultiplexer: isolation, PDL, or transmission slope. While typical production devices may perform better, the values used in this model are typical of, or superior to, specified performance in 2002.

The relationship of actual system performance to the worst case predicted is difficult to guess. If signals in some fibers are cross polarized or close to it, all of the mechanisms with the exception of demultiplexer isolation will be greatly

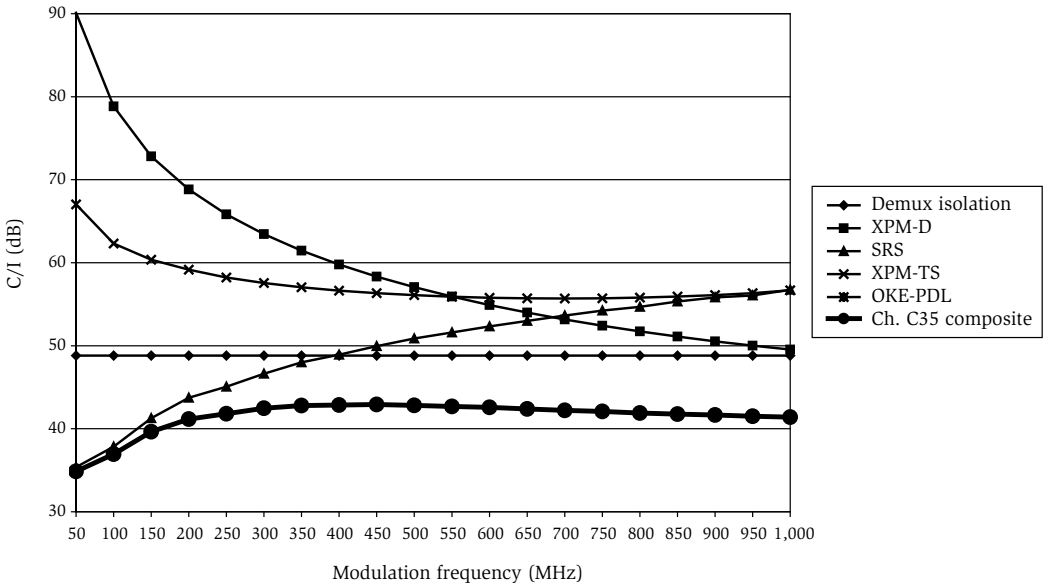


Figure 13.7 Maximum composite C/I: 16 Wavelengths Copolarized, center wavelength.

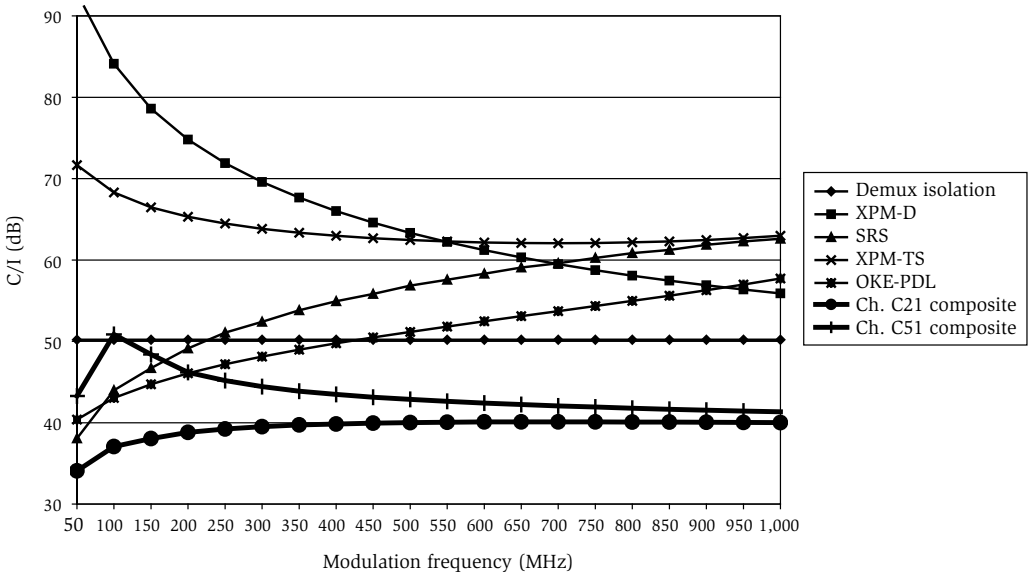


Figure 13.8 Maximum composite C/I: 16 Wavelengths 45° polarized, shortest and longest wavelengths.

reduced. This suggests that the best performance may be close to the 48 to 50-dB limit imposed by the demux isolation, with typical C/I probably falling somewhere in the 42- to 45-dB range for the conditions specified. Whatever the case, it is clear that the various combined cross-modulation effects combine to make such multiwavelength networks inadequate to carry analog video channels.

13.6 CSO Due to Transmitter Chirp Combined with Imperfect Channel Flatness

Directly modulated DFB transmitters exhibit an incidental wavelength modulation, known as *chirp*. Additionally, due to the Kerr effect, each optical carrier generates self-phase modulation (SPM), which adds an additional component of wavelength variation. When transmitted through fiber with appreciable dispersion, these cause second-order distortion in the demodulated signals because the transmission rate through the fiber then varies as a function of the modulation level. Both of these phenomena are covered in Sections 12.6.3 and 12.4.7.

In DWDM systems, however, another factor arises due to the variations in channel amplitude response as a function of optical frequency. Various components may contribute to this, including the optical multiplexer, optical amplifiers, and the demultiplexer. Transmitter chirp interacts with this imperfect channel response to add another stage of optical modulation. Since each modulation is proportional to the modulating current and the modulation stages are effectively in series, the current at the output of the detector will have a component that is proportional to the square of the input current, thus, there will be a second-order distortion mechanism (chirp + transmission slope, or CHIRP-TS). For systems carrying analog video channels exclusively, the calculation is relatively simple.

First, given a change in modulating current Δi_M, the optical power will vary by $\Delta P_{\text{OPT}} = k_1\ \Delta i_M$, where k_1 is the modulation slope of the laser diode. The frequency of the optical carrier will also vary by $\Delta f_{\text{OPT}} = k_2\ \Delta i_M$, where k_2 is the chirp slope of the laser diode in Ghz/mA. The changing optical frequency interacts with the transmission slope of the channel, T_S, in dB/GHz, to produce a change in optical power of $\Delta f_{\text{OPT}} T_S = k_2 \Delta i_M T_S$ dB.

In order to calculate the total effective modulation on the signal as received at the detector, it is necessary to convert this second stage of modulation back into a scaler quantity and to multiply it by the normal laser modulation. When we do that we get

$$\Delta P_R = k_i \Delta i_M 10^{\left(\frac{k_2 \Delta i_M T_S}{10}\right)} \tag{13.13}$$

where

ΔP_R = the change in optical received power resulting from a change in modulating current

k_1 = the modulation slope of the transmitter diode in mW/mA

Δi_M = the change in modulating current in mA

k_2 = the chirp slope of the transmitter diode in Ghz/mA

T_S = the transmission slope of the channel in dB/GHz

Multiplying the change in power by the responsivity of the detector gives us the change in output current as a function of the change in input current and, thus, the equation for the transfer function of the system:

$$\Delta i_O = Rk_i \Delta i_M 10^{\left(\frac{k_2 \Delta i_M T_S}{10}\right)} \tag{13.14}$$

where

R = the detector diode responsivity in mA/mW

This can be simplified by converting the power of 10 to a power of e and then taking the first two terms of the Taylor expansion of $e^x = 1 + x + x^2/2! + x^3/3! \ldots$ This is equivalent to assuming that, over small changes in wavelength, the multiplication of the chirp slope and the transmission slope results in linear modulation as a function of input current. With that simplification, the transfer function becomes

$$\Delta i_O = Rk_i \Delta i_M + Rk_i 0.23026 k_2 T_S \, \Delta i_M^2 \tag{13.15}$$

Referring to the derivation of CSO in Chapter 10, this is in the form of Equation (10.20), but expressed in terms of current rather than voltage, where $i_O = Ai_M + Bi_M^2$, with $A = Rk_1$ and $B = 0.23026 Rk_1 k_2 T_S$. From the analysis in Section 10.3.3, we know that the amplitude of the fundamental is Ai_M and that the amplitude of the individual second-order products is Bi_M^2 and therefore that the ratio of amplitudes of the fundamental and second-order products is A/B, or $1/(0.23026 k_2 T_S I_M)$, where I_M is the peak input modulating current. Since power is proportional to the square of current and the composite second-order product will be determined by the power sum of products, we can write an equation for composite second-order distortion from this mechanism:

$$C/CSO_{CHIRP-TS} = -20 \log(0.23026 k_2 T_S I_M) - 10 \log N \tag{13.16}$$

where

$C/CSO_{CHIRP\text{-}TS}$ = the ratio of the desired carrier on the channel being evaluated to the composite power in the second-order products resulting from the interaction of transmitter chirp and the flatness of the transmission channel

k_2 = the chirp slope of the transmitter diode in Ghz/mA

T_S = the transmission slope of the demultiplexer channel in dB/GHz

I_M = the peak per-channel modulating current in mA

N = the number of second-order beats falling in a cluster affecting the channel being evaluated (from Equation (10.25)).

As was shown in Chapter 12, CSO resulting from the use of directly modulated 1550-nm DFB transmitters combined with transmission through standard fiber resulted in CSO products that made this combination unusable for analog video transmission when the spectrum is greater than an octave. The same is true for transmitter chirp when combined with typical WDM filter transmission slope. For example, a transmission slope of 0.05 dB/GHz (a very moderate assumption because it is typical of just the DWDM demux alone) combined with the typical DFB chirp of 0.25 GHz/mA and 2-mA/channel modulating current results in second-order products that are 44.8 dB below the desired signal. The power addition of the 31 upper-side products affecting the highest channel in a 550-MHz system would result in a C/CSO of only 29.8 dB.

When the RF modulation consists wholly or partially of QAM modulated signals, however, the second-order products are no longer narrow. The low-side product of a QAM channel mixing with an analog video channel exhibits a flat, noiselike spectrum 6 MHz wide centered 1.75 MHz above the lower channel boundary, since the analog carrier is 1.25 MHz above the lower channel boundary and the QAM channel is 3 MHz above the lower channel boundary. The second-order products resulting from the mixture of two QAM channels will be noiselike, centered on the boundary between two channels, but 12 MHz wide (assuming the mixing channels were each 6 MHz wide) with a triangular spectral shape. When a continuous spectrum of adjacent QAM channels is carried, these products will overlap, leading to a relatively flat composite intermodulation noise (CIN) floor whose level can be approximated using the earlier formula, even though the level arises from products each of whose average noiselike density is spread over two channels. For the case of 45 QAM channels extending from 600 to 870 MHz, each modulated at 2 mA/channel, the magnitude of the second-order intermodulation noise per 6 MHz of bandwidth falling near channel 2 would be 29.2 dB below the level of the digital signals. This is 20 dB worse at channel 2 than the CSO, due to chirp interacting with just fiber dispersion (Equation (12.12)) and shows that the degradation due to discrete components often is the limiting factor in system

performance. The import of this severe level of CSO will be discussed later when composite transmission systems are considered.

While this mechanism has been described in the context of the composite optical channel response, it has been investigated by others considering only the response of an optical amplifier, with results that are consistent with the preceding analysis. One experimenter found C/CSO levels as poor as 40 dB for wavelengths between 1535 and 1565 nm but rising to worse than 30 dB at 1525 nm.[14]

13.7 Degradation in Shared-Detector, Multiwavelength Systems

13.7.1 Postdetection White Noise and CIN Addition

When optical signals modulated with nonoverlapping RF spectra are combined and then fed to a common detector, each portion of the detected spectrum will be affected by broadband noise from both transmitters. In the general case, the optical modulation levels of the two transmitters will be different and the relative optical levels and modulation indices will be adjusted at the transmit end so that the levels of the detected signals are in the proper proportion.

As an example, suppose that the signals consist of a spectrum of analog video channels covering the spectrum from 50 to 550 MHz modulating one transmitter and a spectrum of 45 QAM digital signals covering the spectrum from 600 to 870 MHz modulating a second transmitter. It is desired that, after detection, the rms levels of the QAM signal levels be 6 dB below the sync peak levels of the analog video signals.

It is common to set the optical modulation index (OMI) of an externally modulated analog transmitter carrying 78 signals to approximately 3%. This represents a typical compromise among C/N, C/CTB, C/CSO, and clipping probability in the link. It is similarly desired to set the OMI of the QAM transmitter as high as possible, consistent with low clipping probability, to maximize C/N in that portion of the spectrum. When a spectrum consists of many noncoherent signals, the overall peak-to-rms current ratio tends to be independent of the number of signals.[15] Therefore, the probability of clipping remains relatively constant if the total RF drive power remains constant. Since the optical modulation index varies as the drive current, an approximation of the maximum usable per-channel OMI of a transmitter is related to the number of carried channels by

$$m_i = \frac{26.5}{\sqrt{N}} \tag{13.17}$$

where

m_i = the maximum per-channel optical modulation index percentage
N = the number of equal-power RF channels in the modulating spectrum

The constant represents the commonly achievable performance as of the writing of this book. With improved linearization techniques and advanced techniques such as phase control of the modulating carriers and timing control of modulation, it is possible to increase the optical modulation per channel. Given this relationship, the maximum acceptable OMI for the 45 QAM channels in the example is approximately 3.95%.

After detection, each of the recovered RF carriers will be at a level proportional to the square of the OMI, so if the optical carrier levels were the same at the input to the detector, the QAM channels would exceed the analog video channels by $20\log(5.3/3.0) = 2.37$ dB. Since we want those carriers to be lower than the analog carriers by 6 dB, we need to reduce the relative level of the optical signal carrying the QAM channels by about 4.2 dB $[(2.37 + 6)/2]$. It is typical to design analog optical links with a received power of 0 dBm, so the QAM carrier level for proper postdetection level matching would be −4.2 dBm. With the optical system levels now determined, we can calculate the effective C/N for each portion of the spectrum.

In Section 12.10.1 we calculated the C/N of a single optical carrier link. Of the four factors considered, however, only transmitter RIN and IIN are related to each optical signal as received at the detector. The other two, shot noise and postamplifier noise, are related to the detector and the arriving light and so don't add when a common detector and postamp are used.

Thus, the first step is to calculate, independently, just the RIN plus IIN contributions to C/N for the analog (C/N_{A1}) and digital (C/N_{D1}) links. Second, the full link, single-carrier C/N for the analog (C/N_{A2}) and digital (C/N_{D2} — about 2.5 dB better than the analog due to higher OMI, but reduced because of 4.2-dB-lower optical receiver power) should be calculated using Equation (12.21). Third, the C/CIN for the QAM channels due to chirp, combined with fiber dispersion and demultiplexer transmission slope, should be calculated as discussed earlier (C/CIN). Given those, and knowing that the digital signals will be 6 dB below the analog signals after detection, we can calculate the effective carrier-to-thermal noise on the analog link as follows:

$$C/N_{\text{ANALOG}} = -10\log[10^{-C/N_{A2}/10} + 10^{-(C/N_{D1}+6)/10} + 10^{-(C/CIN+6)/10}] \quad (13.18)$$

where

C/N_{ANALOG} = the ratio of analog signal level to thermal noise and noiselike intermodulation noise from the QAM channels

C/CIN = the C/CIN as calculated earlier for channel 2 (the most affected analog channel)

Similarly, the effective carrier-to-thermal noise on the digital link will be

$$C/N_{QAM} = -10\log[10^{-C/N_{D2}/10} + 10^{-(C/N_{A1}-6)/10}] \quad (13.19)$$

A typical CATV downstream network might consist of a broadcast spectrum (50–550 MHz) of analog channels externally modulated (to avoid chirp) and then amplified and split to feed several nodes. The 45 node-specific QAM channels (600–870 MHz) would then each be directly modulated on ITU-grid DFB transmitters and combined in a multiplexer for transmission to a hub. At the hub, they would be demultiplexed; then each would be combined with broadcast programming to be fed to a node. The network designer needs to know the total link C/N for both analog and QAM channels.

If, for example, the fiber cable length is 30 km and the in-channel slope of the demultiplexer is 0.05 dB/GHz, then typical values for the parameters required might be as follows: $C/N_{A1} = 60$ dB, $C/N_{D1} = 65$ dB, $C/N_{A2} = 53$ dB, $C/N_{D2} = 50.5$ dB, $C/CIN_{CHIRP\text{-}DISPERSION} = 49.2$, $C/CIN_{CHIRP\text{-}TS} = 29.4$ (using the assumptions in the earlier example). Using these assumptions, it is not important whether the two C/CIN effects add constructively or destructively, because the CHIRP-TS effect dominates. Plugging these values into Equations (13.18) and (13.19) we get $C/N_{ANALOG} = 35.3$ dB and $C/N_{QAM} = 48.9$ dB. Thus, the degradation from white noise affecting the QAM channels is about 1.6 dB, but it results in a net C/N that is still well above typical end-of-line specifications for this service. On the other hand, clearly this is unacceptable analog performance.

An alternate configuration uses a shared detector at the node but does not use WDM in the QAM circuit before the optical signals carrying the analog and QAM signals are combined. In that case, the noise contribution in the analog spectrum due to second-order intermod products (C/CIN) would be restricted to that associated with fiber dispersion, and net analog C/N would improve to 50.9 dB, which might be acceptable under some circumstances. Given that C/CIN degrades as the square of the length of the path, the performance would likely be acceptable for the relatively short paths between hubs and nodes in headend–hub–node architectures if the combining happened at the hub.

Here are some possible remedies for circumstances where the performance is unacceptable.

- Restrict the QAM spectrum to less than 54 MHz (so that all the second-order products fall below channel 2).
- Give up using a common detector: Use separate detectors at the node, and place a high-pass filter following the QAM detector and a low-pass filter following the analog detector before combining the RF spectra. In the example analyzed, the filter must suppress the second-order intermodulation products at 270 MHz and below by at least 30 dB to avoid significant contribution to total C/N.
- Avoid the use of WDM to transmit multiple QAM spectra ahead of the analog/QAM optical combining.
- Purchase DWDM demultiplexers with flatter response.
- Use externally modulated transmitters for the QAM signals, to eliminate chirp.

13.7.2 Postdetection Discrete Distortion Product Addition

In the usual common-detector configuration, a common spectrum of signals, primarily analog video, uses the lower part of the spectrum and some combination of QAM signals uses the upper part of the spectrum, as discussed earlier. The analog link, considered alone, will include both second-order (CSO) and third-order (CTB) discrete product clusters, which will extend above the analog spectrum. When using a common detector, there is no opportunity to low-pass filter the analog spectrum, with the result that those components will show up as interfering carriers in the QAM spectrum.

Here are the analog components that must be considered:

- Those due to nonlinearity in the optical transmitter/receiver combination
- CSO due to self-phase modulation interacting with fiber chromatic dispersion (discussed in Section 12.4.7)
- If relevant, CSO due to directly modulated DFB transmitter chirp interacting with fiber chromatic dispersion (discussed in Section 12.6.3)
- Clipping distortion products (discussed in Section 12.10.3)
- If relevant, CSO due to directly modulated DFB transmitter chirp interacting with WDM demux in-channel response slope (discussed in Section 13.4.5)

In estimating the probable link performance, it is necessary to consider how the effects add. There is no obvious correlation between residual transmitter/receiver nonlinearities and other effects. The three CSO effects, on the other hand, will either be in phase or directly opposite in phase (depending on the direction of the response slope of the WDM demux). While the probability that clipping will occur is fairly predictable, the magnitude and polarity of individual clipping distortion products are very difficult to predict. The most conservative approach is to assume in-phase addition of all calculable products and thus a "20 log" combination.

Having estimated the worst-channel CTB and CSO products among the analog carriers, Section 10.3.3 discussed the number of products falling in each channel. In the case of CSO products, Equation (10.25) can be used to calculate the number of products in each channel, including above the analog spectrum, and it can be assumed that the number of third-order products will be greatest near the middle of the analog spectrum and that the greatest magnitude of CTB distortion will fall higher in the spectrum. Knowing the worst-case distortion and the number of distortion products as a function of frequency, the magnitude of C/CTB and C/CSO distortion in any channel can be calculated by assuming that the distortion magnitude will vary as "10 log n," where n is the number of products in that channel. Finally, the ratio of digital carrier to analog C/CSO or C/CTB products will be lower (worse) than the foregoing, by the ratio of analog signal levels to QAM signal levels in the combined RF spectrum (typically 6 dB).

Only the upper-side CSO products (2.5 MHz above the lower channel boundary) need be considered because only those products extend above the analog spectrum and because the lower-side products fall between standard channel boundaries. CTB products will fall 1.25 MHz above the lower channel boundary.

If we look at the previous example of a 30-km optical link and assume that the analog signals, occupying 50–550 MHz, modulate an externally modulated transmitter operating in the 1550-nm range with a launch power of +17 dBm, we get the following contributions to distortion:

- The transmitter/receiver pair will typically generate upper-side C/CSO and C/CTB of 65 dB on the most affected channels.
- Self-phase modulation will interact with the fiber to produce C/CSO at a level of about 70 dB.

At worst, the CSO effects will add on a voltage basis to produce a total C/CSO of 61 dB at the highest analog channel. Assuming the digital spectrum occupies 600–870 MHz, the level at the lowest digital channel will be 3.7 dB lower relative to the analog carriers (from Equation (10.25) and assuming that CSO varies as $10\log(n)$, where n is the number of high-side products) but will be higher relative to the QAM signals that are run at 6-dB-lower levels than analog, giving a net C/CSO in the lowest QAM channel of 58.7 dB.

The analog CTB will be highest somewhere above the center of the analog spectrum and fall off by approximately 2 dB at the upper edge of the analog spectrum and by about 3 dB at the lowest QAM channel. As with CSO, however, we need to account for the 6-dB difference in levels between analog and QAM signals, giving us a net C/CTB for the lowest QAM channel of about 62 dB.

Given a typical end-of-line C/(N + I) spec of 40 dB, the levels of CTB and CSO products resulting from use of a common detector are seldom a problem, provided externally modulated analog transmitters are used.

13.8 Summary of WDM Link Performance

Wavelength-division multiplexing has become standard in the engineering of cable television and similar networks because it facilitates the delivery of switched services to small groups of customers. It does this by allowing the transport of many independent signals over shared fibers and through shared optical amplifiers.

Unfortunately, optical signals on separate wavelengths interact as they travel through the fiber, and those interactions, sometimes in conjunction with discrete optical components in the circuit, generate various levels of crosstalk. Depending on the parameters of a given link, these mechanisms can have a serious effect on recovered RF signal quality. As our analysis shows, the quality of discrete components in general, and of the wavelength demultiplexer in

particular, is typically the limiting factor in achieving acceptable low levels of crosstalk interference.

The analysis also shows that the achievable quality for 16-wavelength DWDM circuits with lengths approaching 30 km will not be acceptable for analog video but will be adequate for at least 256 QAM modulated signals with adequate component quality.

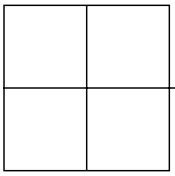
Endnotes

1. Bernard Eichenbaum, *Coarse WDM Applications, Architectures, and Scalability for HFC and FTTH Digital Broadband Access*, Proceedings Manual and Collected Technical Papers, Cable-Tec 2002 Expo, SCTE, Exton, PA, June 2002, pp. 19–38.
2. New Global Standard Set for Metro Networks, *IEEE Spectrum*, August 2002, pp. 21, 24.
3. For a brief, nonmathematical summary comparing wavelength multiplexer technologies, see Schlomo Ovadia, *Broadband Cable TV Access Technologies*. Prentice-Hall, Upper Saddle River, NJ, 2001, Sec. 6.1.1.
4. Based on data sheets for current products from Scientific Atlanta and Arris. (both located in Atlanta, Georgia).
5. Much of the material in this section related to XPM, OKE-PDL, and SRS is based on the excellent and clearly explained work of M. R. Phillips (sometimes copublished with D. M. Ott), various of whose papers covering optical cross modulation are referenced in subsequent endnotes.
6. K. Y. Wong et al., Nonlinear Crosstalk Suppression in a WDM Analog Fiber System by Complementary Modulation of Twin Carriers, OFC2001 conference, Optical Society of America, 2000, pp. WV5-1 – WV5-3.
7. M. R. Phillips and D. M. Ott, Crosstalk Caused by Nonideal Output Filters in WDM Lightwave Systems. *IEEE Photonics Technology Letters*, Vol. 12, No. 8, August 2000.
8. Mary R. Phillips and Daniel M. Ott, Crosstalk Due to Optical Fiber Nonlinearities in WDM CATV Lightwave Systems. *Journal of Lightwave Technology*, Vol. 17, No. 10, October 1999.
9. Ibid, p. 1788.
10. Ibid, p. 1790.
11. Kyo Inoue, Four-Wave Mixing in an Optical Fiber in the Zero-Dispersion Wavelength Region. *Journal of Lightwave Technology*, Vol. 10, No. 11, IEEE, November 1992, pp. 1553–1561.
12. James J. Refi, Optical Fibers for Optical Networking, *Bell Labs Technical Journal*, January–March 1999, pp. 246–261.

13. M. R. Phillips, Crosstalk in an Eight-Wavelength WDM Analog Lightwave System: Measurement and Analysis. Optical Fiber Communication Conference and Exhibit, 2001, *OFC 2001*, Vol. 3, pp. WCC1, 1–3.

14. C. Y. Kuo and E. E. Bergmann, Erbium-Doped Fiber Amplifier Second-Order Distortion in Analog Links and Electronic Compensation, *IEEE Photonics Technology Letters*, Vol. 3, No. 9, September 1991, pp. 829–831.

15. Section 17.3 of the *Recommended Practices for Measurements on Cable Television Systems*, NCTA, 2002, contains an excellent tutorial on this subject. It shows that the probability of exceeding a given peak-to-rms voltage ratio is essentially unchanged as the number of summed signals increases beyond 10.



Chapter 14
Linear Microwave Signal Transportation

14.1 Introduction

Although linear amplitude modulated fiber-optic links are the favored trunking methodology for most applications, there are situations where they are simply not cost-effective. For example, physical barriers, such as waterways, may make cable construction very expensive. Alternatively, the network may need to serve several small outlying communities that are so widely separated that the cost of the interconnecting links affects the economic viability of the project. Finally, network operators sometimes need to transport signals through areas where rights-of-way are either unavailable or very expensive to acquire. In any of these cases, point-to-point or point-to-multipoint broadband amplitude modulated microwave links (AMLs) may offer a superior solution.

This chapter will cover the basic operation of this equipment and the essential calculations and methodology required to engineer AMLs. No attempt is made to present a comprehensive manual on either internal designs of microwave equipment or the many subtleties of path engineering. Several standard reference books on the subject are available for those desiring to pursue the subject in greater depth.[1–4]

Microwave may be used for relay of individual channels or for transport of the entire FDM spectrum. The use of channelized microwave (which often employs digital or FM modulation) is covered in Chapter 7. The discussion here will be limited to links that transport the entire spectrum and are part of the broadband distribution network.

14.2 U.S. Regulation of Microwave Transmission

Regulation of radio transmission will vary by country, though broad spectrum planning is harmonized worldwide through international conferences. In the United States the essential regulations pertaining to AML systems are contained in the FCC's rules.[5] In particular, Part 17 covers towers, whereas Part 78 covers microwave licensing, channelization, and performance. Operators of microwave systems are required to have current copies of the rules on hand and to comply

with detailed requirements regarding licensing, log books, posting of stations, periodic performance tests, and the like. Maintenance of tower lighting, where required, is a matter of critical concern because of aircraft safety issues.

The FCC has set aside the frequencies extending from 12.7 to 13.2 GHz for Cable antenna relay service (CARS). Some additional spectrum is assigned at 18 GHz and 31 GHz, as well. Within the 12-GHz band, several acceptable channelization schemes are given for both FM and AM modulation: A, B, and K channels for FM and C, D, E, and F designations for AML. Each channel in each scheme is given a unique identifier.

In the case of the AML bands, each channelization group represents a direct translation of the FDM cable spectrum to a different portion of the assigned microwave band as follows:

- Group C maps 54 through 300 MHz to 12,700.5 through 12,946.5 MHz.
- Group D maps 54 through 300 MHz to 12,759.7 through 13,005.7 MHz.
- Group E maps 54 through 300 MHz to 12,952.5 through 13,198.5 MHz.
- Group F maps 54 through 240 MHz to 13,012.5 through 13,198.5 MHz.

Transmission of a full 80-channel spectrum is possible through the use of a combination of Groups C, D, and E channels (with some minor shifts from nominal microwave frequencies) occupying the full 500-MHz CARS band. Clearly, without some form of frequency reuse, no combination of available 12-GHz channels is adequate for systems whose bandwidth exceeds 500 MHz.

In general, the FCC limits the microwave power level delivered to the antenna system to 5 watts per channel and the radiated power to +55 dBW EIRP.* Microwave frequencies must be accurate within ±0.005%, and the original cable channel frequencies must not shift by more than ±0.0005% through the link.

14.3 General Operational Principals

The basic principal of operation of an AML link is simplicity itself. At the transmitter, the incoming cable spectrum is simply mixed with an unmodulated signal (the microwave local oscillator, or LO) whose frequency is equal to the difference between the original channel frequencies and the translated microwave frequencies. At the receiver, the microwave spectrum is mixed with another unmodulated carrier at the same LO frequency, resulting in recovery of the original cable spectrum. In order to preserve the frequency accuracy of the

* See Chapter 7 for a discussion of the relationship among antenna gain, transmitter power, and effective isotropic radiated power (EIRP).

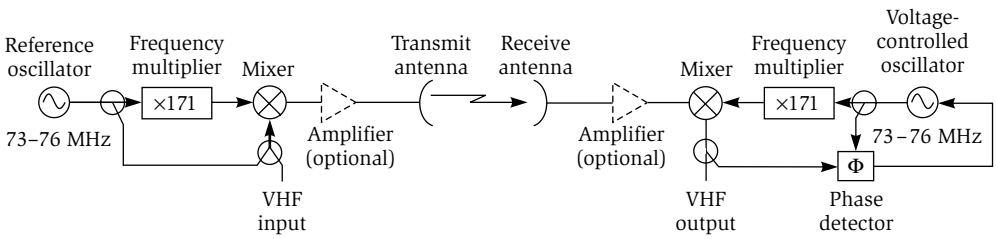


Figure 14.1 Basic AML principles of operation.

VHF spectrum, as well as provide the required microwave frequency accuracy, a crystal reference oscillator operating at 1/171 of the offset frequency is used to phase lock the transmitter local oscillator directly. A sample of the reference frequency (which falls in the VHF spectrum between cable channels 4 and 5) is also up-converted and sent along with the remainder of the cable spectrum through the microwave link. At the receiver, the reference signal is recovered and used to phase lock the microwave LO there. Since both LOs are locked to a common reference, the recovered VHF spectrum is frequency coherent with the transmitted spectrum, while the accuracy of the microwave spectrum depends on the accuracy of the reference oscillator. Figure 14.1 is a basic functional diagram of an AML. Table 14.1 summarizes the frequencies for various bands, and Table 14.2 shows the combination of AML channels used to transport a 54 to 552-MHz cable spectrum.

Transmitter designs vary in the way channels are handled and in their power capabilities. Some early AML transmitters used a separate transmitter for each cable channel; most modern units are broadband. High-power transmitters use microwave amplification; low-power units may use the mixer output directly. Finally, early units used klystron tubes; most modern units use solid-state

Table 14.1 Frequency Relationships in AML Links (in MHz)

FCC channel group	*Cable spectrum*		*Microwave spectrum*		*Offset frequency*	*Reference oscillator*
	Start	*Stop*	*Start*	*Stop*		
C	54	552	12,700.5	13,198.5	12,646.5	73.9566140
D	54	300	12,759.7	13,005.7	12,705.7	74.302339
E	54	300	12,952.5	13,198.5	12,898.5	75.429824
F	54	240	13,012.5	13,198.5	12,958.5	75.780701

Table 14.2 AML Channel Combination Used to Transmit the Cable Spectrum from 54 to 552 MHz

AML channel	*Cable channel*	*Channel limits (MHz)*		*AML channel*	*Cable channel*	*Channel limits (MHz)*	
		Low	*High*			*Low*	*High*
C01	2	54	60	D33	37	300	306
C02	3	60	66	E01	38	306	312
C03	4	66	72	E02	39	312	318
C04	AML pilot	72	76	E03	40	318	324
C05	5	76	82	E04	41	324	330
C06	6	82	88	E05	42	330	336
C07	FM	88	94	E06	43	336	342
C08	FM	94	100	E07	44	342	348
C09	FM	100	106	E08	45	348	354
C10	FM	106	108	E09	46	354	360
C11	98	108	114	E11	47	360	366
C12	99	114	120	E12	48	366	372
C13	14	120	126	E13	49	372	378
C14	15	126	132	E14	50	378	384
C15	16	132	138	E15	51	384	390
C16	17	138	144	E16	52	390	396
C17	18	144	150	E17	53	396	402
C18	19	150	156	E18	54	402	408
C19	20	156	162	E19	55	408	414
C20	21	162	168	E20	56	414	420
C21	22	168	174	E21	57	420	426
C22	7	174	180	E22	58	426	432
C23	8	180	186	E23	59	432	438
C24	9	186	192	E24	60	438	444
C25	10	192	198	E25	61	444	450
C26	11	198	204	E26	62	450	456
C27	12	204	210	E27	63	456	462
C28	13	210	216	E28	64	462	468
C29	23	216	222	E29	65	468	474
C30	24	222	228	E30	66	474	480
C31	25	228	234	E31	67	480	486
C32	26	234	240	E32	68	486	492
C33	27	240	246	E33	69	492	498
C34	28	246	252	E34	70	498	504
C35	29	252	258	E35	71	504	510
C36	30	258	264	E36	72	510	516
C37	31	264	270	E37	73	516	522
C38	32	270	276	E38	74	522	528
C39	33	276	282	E39	75	528	534
C40	34	282	288	E40	76	534	540
C41	35	288	294	E41	77	540	546
C42	36	294	300	E42	78	546	552

Note: Channel D33 offset +0.8 MHz from normal; channels E04–E09 offset +2.0 MHz from normal.

components, though the highest-powered units still use traveling wave tubes (TWTs), which provide high power with both gain and phase stability.

Regardless of the difference in transmitters, all receivers are similar, differing primarily in their mounting, powering, and flexibility in connecting to the downstream coaxial distribution system. In some cases, microwave low-noise amplifiers (LNAs) are used to improve the noise figure of receivers and/or allow the receiver to be mounted at the base of a tower rather than at the back of the antenna.

14.4 Path Design

Microwaves travel approximately along "line-of-sight" paths, meaning that, unlike low-frequency AM broadcast stations, they do not bend significantly to follow the earth's curvature (though a slight bending in the direction of the earth is typical). Like light waves, however, they can be reflected from multiple intermediate surfaces between the transmitter and receiver. Under some atmospheric conditions, microwave signals can temporarily shift from their normal path, either toward or away from the earth's surface. Such bending, discussed in Section 14.6, can cause the transmitted signal to completely miss the receiving antenna and is one of the factors limiting link reliability.

Designing a microwave path between two fixed points consists of determining the distance and direction between the points and then examining the path to make sure there is adequate clearance from any objects near the direct line between transmitting and receiving antennas.

14.4.1 Path End Coordinates

The locations of the path ends are described by four parameters: the longitude and latitude, which define the location on the earth's surface, and the altitude of the base of the antenna tower and the mounting height of the antenna on that tower, which together define the vertical location of the antenna relative to sea level.

The longitude is usually expressed as the sum of the number of degrees, minutes, and seconds that the location is east (negative) or west (positive) of the prime meridian, which passes through Greenwich, England. The latitude is the sum of the number of degrees, minutes, and seconds that the location is north (positive) or south (negative) of the equator.

Since there are 60 minutes in a degree of arc and 60 seconds in a minute of arc, we can convert the longitude and latitude of the transmitter and receiver antenna locations to decimal equivalents as follows:

$$\text{longitude or latitude (decimal degrees)} = \text{degrees} + \frac{\text{minutes}}{60} + \frac{\text{seconds}}{3{,}600} \quad (14.1)$$

For the remainder of this chapter, we will assume that this conversion has been done and that the transmitting location is at a west longitude of W_T and a north latitude of N_T, while the receiving antenna is at a west longitude of W_R and a north latitude of N_R, where all numbers are in decimal degrees. Minor modifications of the formulas will be required for links located in the southern hemisphere and/or eastern longitudes.

14.4.2 Path Length

Path calculations are based on an average earth circumference of 24,857 statute miles. Thus, each degree of arc along that surface represents $24{,}857/360 = 69.047$ statute miles.

We can approximate the path length by calculating the length of the diagonal of a hypothetical rectangle whose corners are located at the transmit and receiver locations, as shown in Figure 14.2. In the north–south direction, the distance in miles is just the difference in the latitude values times 69.047, because latitude lines are all the same length. In the east–west direction, however, one degree represents that distance only at the equator. At any distance away from the equator, the distance is shortened by the cosine of the latitude. Therefore the width of the rectangle in miles is approximately the difference in longitudes times the cosine of the average of the latitudes times 69.047.

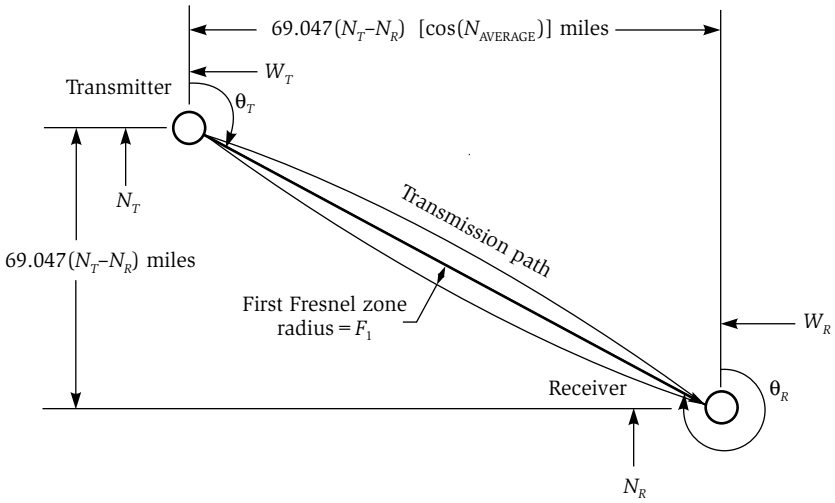


Figure 14.2 Microwave path horizontal parameters.

The path distance, D, is the length of the diagonal of this rectangle, calculated by taking the square root of the sum of the squares of the height and width of the rectangle, or*

$$D \text{ (statute miles)} \simeq 69.047\sqrt{(N_T - N_R)^2 + \left[(W_T - W_R)\cos\left(\frac{N_T + N_R}{2}\right)\right]^2} \quad (14.2)$$

Though this gives close results for the path lengths likely to be involved in most AML applications, the precise formula for path length is

$$\begin{aligned} D \text{ (statute miles)} = 69.047 \arccos[&\sin(N_T)\sin(N_R) \\ &+ \cos(N_T)\cos(N_R)\cos(W_T - W_R)] \end{aligned} \quad (14.3)$$

If the path length is desired in nautical miles for some reason, then use 60 rather than 69.047 to convert degrees of arc to distance (nautical miles are based on an earth circumference of exactly 21,600 miles).

14.4.3 Path Azimuth

The next required value is the compass heading (azimuth) for the antennas, in order to align the path. Compass headings are measured in degrees clockwise from true north. The first step is to calculate the angle θ using

$$\theta(\text{degrees}) = \arccos\left[\frac{\sin(N_R) - \sin(N_T)\cos(D/69.047)}{\sin(D/69.047)\cos(N_T)}\right] \quad (14.4)$$

If W_R is less than W_T, then the transmit antenna azimuth $\theta_T = \theta$. If W_R is greater than W_T, then $\theta_T = 360^\circ - \theta$.

The azimuth for the antenna at the receive site is the opposite of that at the transmit site, that is, for θ_T between 0° and 180°, $\theta_R = \theta_T + 180$ degrees, whereas for θ_T between 180° and 360°, $\theta_R = \theta_T - 180^\circ$.

14.4.4 Path Clearances

Having determined the horizontal distance and azimuth, the next step is to check for adequate clearances along the path. In the horizontal direction, this requires a calculation of a factor known as the *Fresnel zone*; in the vertical direction, it additionally requires consideration of the topology of the land, the height of such features as trees and buildings, and the effective curvature of the earth.

---

* Unless otherwise specified, the arguments for trigonometric functions are in degrees, not radians.

Fresnel Zone Calculations

The primary path from transmitting to receiving antenna is a nominally straight line. It is possible, however, for the signal to be reflected from objects that, though not in the direct path, are close enough that they are within the beamwidth of the antennas. If that happens, whatever portion of the transmitted signal is reflected toward the receiving antenna will arrive there after the direct signal because of the longer path length involved. The impact of this reflected wave will depend on the nature of the surface from which the signal is reflected, on the difference in path length, and on the pattern of the transmitting and receiving antennas.

The first Fresnel zone is defined as an imaginary ring surrounding the center of the direct path such that the distance from the transmitting antenna to this ring plus the distance from the ring to the receiving antenna is equal to one-half wavelength more than the direct path between the antennas. Subsequent Fresnel zones are defined as larger imaginary rings, where the difference in path lengths is $n\lambda/2$, where n is an integer larger than 1 and λ is the free-space wavelength of the signal.

The radius of the first Fresnel zone at a point D_1 miles from the transmitting antenna is

$$F_1 = 72.1\sqrt{\frac{D_1 D_2}{fD}} \tag{14.5}$$

where

F_1 = the radius of the first Fresnel zone in feet
D = the total direct path length in miles
D_1 = the distance from the transmitting antenna in miles
D_2 = the distance from the receiving antenna in miles $= D - D_1$
f = the operating frequency in GHz

Figure 14.2 shows how the width of the first Fresnel zone varies along the transmission path.

Although higher-order Fresnel reflections can cause problems with highly reflective surfaces that are oriented optimally (such as smooth, flat terrain or calm water), the general rule of microwave design is that it is sufficient to clear objects by $0.6F_1$.

Additional Vertical Clearances

In the vertical plane, it is necessary to add the Fresnel zone clearance to the effective height of objects that lie under the direct path. Factors to be considered include the curvature of the earth, peaks in the terrain, and/or any objects, such as buildings and trees. Traditionally, such analyses were done graphically, although they are more likely to be done with a computer today. Figure 14.3 illustrates how the terms add up.

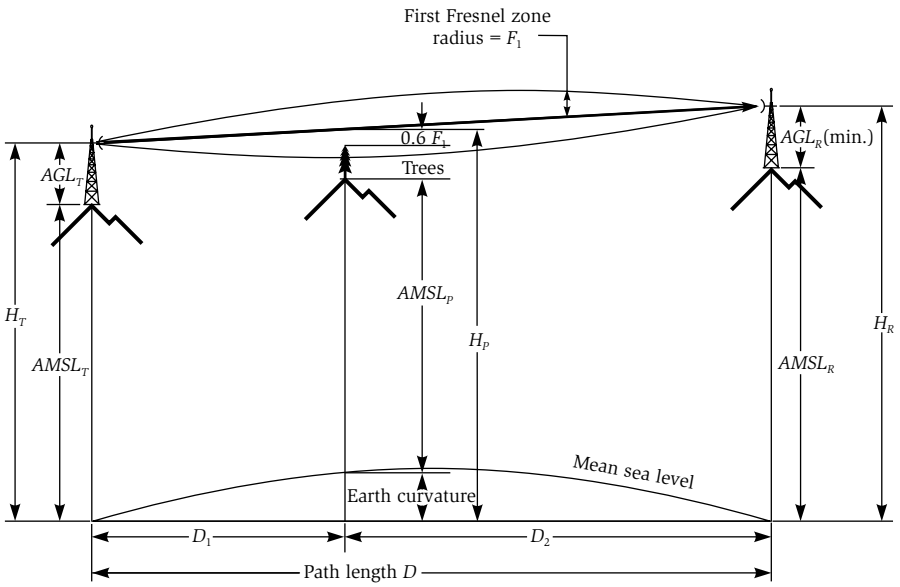


Figure 14.3 Microwave path vertical clearances.

The flat line at the bottom of the Figure represents a straight line drawn through the earth from mean sea level at the transmit location to mean sea level at the receive location. At the transmit end, a vertical line represents the elevation of the ground at the base of the antenna tower, as surveyed or shown on a topological map. The units are feet *above mean sea level* (*AMSL*), we have designated this elevation as $AMSL_T$ to identify which site is referenced. The elevation of the center of the transmit antenna relative to the ground is *above ground level*, or AGL_T. The sum of the ground height and antenna mounting height we have designated as H_T. At the other end of the path, the equivalent values are $AMSL_R$, AGL_R, and H_R.

The curved line that intersects the ends of the flat line represents the effective curvature of the earth between the endpoints. The effective curvature may not equal the actual curvature, however, because the microwave beam may be slightly bent due to atmospheric conditions. While the beam most commonly bends toward the earth (thereby allowing greater clearance over obstacles in the path), occasionally the reverse is true. These effects are taken into account using a factor called the *K* factor. *K* is the ratio between the effective and actual earth radius. The effect of earth curvature along a path of length *D* miles can be calculated using

$$h = \frac{D_1 D_2}{1.5\,K} \tag{14.6}$$

where

h = the virtual height, in feet, due to effective earth curvature
D_1 = the distance to the transmitting antenna in statute miles
D_2 = the distance to the receiving antenna in statute miles = $D - D_1$
K = the K factor

Under normal atmospheric conditions, K is approximately 4/3. Conservative path design, however, frequently calls for calculating clearances using $K = 2/3$. K for any specific location is determined from a factor known as the A*ugust mean radio refractivity* and the path elevation above mean sea level.

In the special case of equal-height antennas transmitting over a smooth surface (so that the maximum interference point is midway along the path), Equation (14.6) reduces to

$$h_{\max} = \frac{D^2}{6K} \tag{14.7}$$

where

$h_{\max}$ = the effective height increase, in feet, at the midpoint of the path
D = the total path length in miles

The next step is to examine the vertical path profile. Generally this is done using topological maps, such as those available from the United States Geological Survey (USGS). Rather than plot the entire profile, it is sufficient to plot selected points that represent elevation maxima. For simplicity, we have shown a single peak, located at distance D_1 from the transmit antenna, whose effective height is $AMSL_P$ plus the effective curvature of the earth at that distance along the path.

Added to this effective peak height must be anything that projects above the ground level at the peak. In wooded areas, for instance, it is common to add a number of feet that represents the maximum height of trees in that area; in a more urban setting, it might be a specific building. Finally, we must add $0.6F_1$, the required clearance to the first Fresnel zone, to get H_P, the total effective height of the peak.

Using the height of the transmit antenna, the effective height of the peak, and its distance from the transmit antenna, we can calculate the minimum mounting height for the receive antenna using

$$AGL_R(\text{ft}) \geq (H_P - H_T)\left(\frac{D}{D_1}\right) + H_T - AMSL_R \tag{14.8}$$

So long as the receive antenna mounting height above ground is greater than this value, the path should be acceptable from the standpoint of vertical clearances.

Note that in the special case of a flat landscape covering much of the distance between antennas, it may not be obvious where along the path the greatest potential interference occurs. A reasonable approximation requires first calculating the effective difference, ΔH, in elevation between the transmit antenna and the flat ground along the path:

$$\Delta H(\text{ft}) \simeq AMSL_T + AGL_T - AMSL_P - \text{Trees} - 21.63\sqrt{\frac{D}{f}} \tag{14.9}$$

where

$AMSL_T$ = the elevation at the base of the transmit tower
AGL_T = the mounting height on the tower
$AMSL_P$ = the elevation of the flat land between the antennas
Trees = the allowance for foliage and objects on the flat land
D = the total distance between the antennas in miles
f = the operating frequency in GHz

The only approximation in Equation (14.9) is the last term, which calculates the Fresnel clearance on the assumption that the highest point will be roughly in the center of the path. Obviously, to the degree that the total heights of the transmitting antenna and of the receiving antenna are different, this will be in error.

Given the effective difference in elevation between the transmit antenna and the flat area, the maximum distance that a signal can travel from the transmitter and still clear the ground is

$$D_1 \simeq 69.047 \arccos\left(\frac{R_E K}{R_E K + \Delta H}\right) \tag{14.10}$$

where

D_1 = the distance from the transmitter, in miles, to the point where the signal just clears the ground by the required amount
R_E = the radius of the earth in feet = 20,888,284
K = the K factor
ΔH = the effective elevation difference calculated earlier

D_1 can now be plugged into Equation (14.8), along with the other factors, to determine the required mounting height of the receive antenna.

14.5 Performance Calculation

In order to predict the performance of an AML, we first need to define the details of the complete signal path from the transmitter to each receiver. Then we will start with the transmitter output RF power per channel and add up all the signal losses and gains (in dB) from the path elements to determine the nominal power level at each significant point. Knowing the performance of each circuit element, we will then calculate the cascaded noise and distortion of the link under normal weather conditions. Finally, using charts of expected weather disturbances, we will estimate the reliability of the circuit.

14.5.1 Signal Path Definition

The first step in performance calculation is to carefully list every circuit element in the path between transmitter and receiver. Since a single transmitter may feed more than one receiver, the first element may well be a signal splitter. Following that will usually be sections of waveguide. Since elliptical waveguide is more lossy than circular, it is not uncommon to use circular waveguide to feed up tall towers and then to transition to elliptical guide for the last few feet at both ends of the transmission line to allow flexibility for dish adjustments and equipment placement.

The dish will have a defined gain, followed by the path loss through the air, followed by the gain of the receiving antenna. Finally, there may be a low-noise amplifier at the output of the receive antenna, followed by some combination of waveguides down the tower and finally terminating at the receiver input. All the applicable circuit elements should be listed separately for each section of the path and a separate list made up for each path fed from a transmitter. The left two columns of Table 14.3 show a sample list of circuit elements for a typical AML path at 13 GHz.

14.5.2 Power Budget Calculation

In the first row of the fourth column of the worksheet is the transmitter power output per channel in dBm. Unlike coaxial cable distribution networks, where the reference for power measurements is 1 millivolt in a 75-ohm impedance (dBmV), the reference for microwave power is the milliwatt $= 10^{-3}$ watts. Thus the power in dBm $=$ $10 \log (P)$, where P is the power in milliwatts. AML transmitters are available with rated power output levels per channel varying from about -12 dBm to $+33$ dBm.

Next come any splitters used to feed multiple receive sites. Manufacturer's data sheets will give the loss of these. As a first approximation, the loss will typically be a few tenths of a decibel higher than the theoretical splitter ratio loss (e.g., a four-way splitter has a theoretical loss of 6 dB and an actual loss of about 6.5 dB to each port). Enter the appropriate loss in the third column of the worksheet as a negative number (representing "negative gain" in that path element).

Table 14.3 Sample AML Performance Calculation

Network element	*Quantity/ property*	*Gain (dB)*	*Cumulative power (dBm/ch)*	*C/N (dB)*	*CTB (dBc)*	*CSO (dBc)*
Transmitter			+10.0	61.0	−65.0	−70.0
Splitter	2-way	−3.5	+6.5			
Circular waveguide	200 ft	−2.8	+3.7			
Elliptical waveguide	50 ft	−1.9	+1.8			
Transmit antenna	10 ft dia.	+48.8	+50.6			
Free-space loss	24 mi	−146.5	−95.9			
Excess loss allowance		−2.0	−97.9			
Receiving antenna	10 ft dia.	+48.8	−49.1			
LNA	3 dB NF	+17.0	−32.1	55.9		
Circular waveguide	300 ft	−4.1	−36.2			
Elliptical waveguide	50 ft	−1.9	−38.1			
Receiver	13 dB NF			56.9	−70.0	−61.0
Total Link Performance				52.6	−61.1	−58.4

After the splitter will be the waveguides connecting the splitter output port to the antenna. As a general rule, circular waveguide has a loss of about 0.014 dB/ft at 13 GHz, while elliptical waveguide has a loss of about 0.038 dB/ft at the same frequency. The losses of transmission lines used at other frequencies will be given in the manufacturer's literature or in commonly available design tables. For each type of transmission line, multiply the line length by the loss per unit length and enter the net loss into the third column, again as a negative gain number.

Next enter the transmitting antenna gain, which will be positive relative to an isotropic radiator. Parabolic antenna design and performance are treated in detail in Chapter 8. The antennas used for CARS band are almost always constructed with prime focus feeds. Diameters commonly range from 4 to 10 feet. While actual antennas will vary slightly, a typical gain for a circular antenna with a parabolic cross section and a prime focus "button hook" feed is

$$\text{parabolic antenna gain (dBi)} = 20\log(d) + 20\log(f) + 6.5 \qquad (14.11)$$

where

d = the diameter of the antenna in feet
f = the operating frequency in GHz

The loss suffered as a result of sending the signal through the air, the *free-space loss*, is dependent on both path length and operating frequency. Under normal atmospheric conditions, free-space loss is

$$\text{path loss (dB)} = 96.6 + 20\log(f) + 20\log(D) \tag{14.12}$$

where

f = the operating frequency in GHz
D = the path distance in statute miles

Next, it is common to add an additional loss allowance of about 2 dB to account for aging, imperfect alignment of equipment, and so on (sometimes known as the *field factor*).

The receive antenna will have a gain that will also be calculated in accordance with Equation (14.11). If an LNA is mounted at the back of the receive antenna feedhorn, as is common, enter its gain at this point. Now, as at the transmitter end, enter all the feedline losses between the antenna or LNA and receiver input.

In the fourth column, calculate the power at each point in the circuit, adding power gain values and subtracting losses in dB. The final number is the receiver input level, in dBm, under normal atmospheric conditions.

14.5.3 Carrier-to-Noise Calculation

As discussed in chapter 11, the room-temperature thermal noise power in a 4-MHz bandwidth is about −59 dBmV. 0 dBmV, however, is approximately equal to −49 dBm, so the noise level referenced to 1 mW is −108 dBm.

Thus, the equivalent input noise floor of any device is equal to $-108\,\text{dBm} + F_A$, where F_A is its noise Figure. Its C/N contribution (the difference between the driving signal level and the input noise floor) can be calculated using

$$C/N_{\text{Device}}(\text{dB}) = 108 - F_A(\text{dB}) + P_{\text{IN}}(\text{dBm}) \tag{14.13}$$

where

C/N_{Device} = the carrier-to-noise ratio contribution of the device
F_A = the noise Figure of the device in dB
P_{IN} = the driving power level in dBm.

There are three principal contributors to link C/N: the transmitter, the LNA (if used), and the receiver. The transmitter C/N is generally given by the manufacturer for various channel-loading conditions. Since the input levels and noise Figures of both LNA and receiver are known, their C/N contributions are readily calculable. All three C/N values are entered into column five of the worksheet.

The link C/N is calculated using the methodology of Equation (11.2):

$$C/N_{ttl} = -10\log\left(10^{-\left(\frac{C/N_T}{10}\right)} + 10^{-\left(\frac{C/N_L}{10}\right)} + 10^{-\left(\frac{C/N_R}{10}\right)}\right) \tag{14.14}$$

where C/N_{ttl} is the carrier-to-noise ratio of the entire link and C/N_T, C/N_L, and C/N_R are the carrier-to-noise contributions of the transmitter, LNA, and receiver, respectively.

14.5.4 Distortion Calculation

The CTB and CSO distortion levels are given by the manufacturers for various channel loadings and power levels. In most cases, the LNA will contribute only slightly to the overall distortion levels.

Addition of second- and third-order distortion levels is discussed in chapter 11. The link performance numbers in the worksheet were calculated using the most conservative assumption, which is

$$C/CTB_{\text{Link}} = -20\log\left(10^{-\left(\frac{C/CTB_T}{20}\right)} + 10^{-\left(\frac{C/CTB_R}{20}\right)}\right) \tag{14.15}$$

where

C/CTB_T = the transmitter CTB in dB
C/CTB_R = the receiver CTB in dB

A similar formula is used to calculate link CSO performance. Manufacturers should be consulted for the specific microwave equipment chosen to determine whether Equation (14.15) is appropriate for calculating cascaded distortion.

14.6 Link Availability Factors

14.6.1 Multipath

The preceding calculation predicts the performance of the microwave link under normal weather conditions. Two significant atmospheric effects can degrade the signal, however: multipath and rain fade. An understanding of the general causes and methods for calculating their effects on link performance is important.

It will be recalled that generally a microwave beam bends slightly toward the earth (with the result that the earth looks "flatter" than it actually is). The principal cause of this bending is that signals are not transmitted through a vacuum, but rather through the air. Since the density of air usually decreases with altitude, there will be a slight gradient across the width of the microwave beam, causing a slight decrease in velocity of propagation at the bottom of the beam. This causes the beam to bend slowly toward the earth.

Unfortunately, the density and change in density of the air is not a constant, with the result that sometimes the beam bends more and sometimes less. Under extreme conditions, the beam can miss the receiving antenna altogether. Under other conditions, there may also be more than one path between transmitter and receiver, with slightly different transit times and, therefore, phase and amplitude variations at the receiver. These effects are called *multipath*.

Since multipath degradation is a statistically random effect, the usual way of quantifying its effect is to define a minimum usable condition for the link and then to predict how much of the time the link performance exceeds this condition (defined as the *availability* of the path). For analog video links, a common definition of the minimum usable C/N is 35 dB.

The first step in calculating path availability is to determine how much the path loss can deteriorate before the C/N reaches 35 dB (the *fade margin*). Realizing that the C/N of both LNA and receiver will degrade linearly with increasing path loss while the transmitter's contribution will stay constant, we can write a single equation for fade margin:

$$M(\text{dB}) = 10\log\left(10^{-\left(\frac{C/N_{\min}}{10}\right)} - 10^{-\left(\frac{C/N_T}{10}\right)}\right) - 10\log\left(10^{-\left(\frac{C/N_L}{10}\right)} + 10^{-\left(\frac{C/N_R}{10}\right)}\right) \quad (14.16)$$

where

M = the fade margin in dB
$C/N_{\min}$ = the defined minimum usable link C/N, usually 35 dB
C/N_T = the transmitter's C/N
C/N_L = the LNA C/N under nominal signal conditions
C/N_R = the receiver C/N under nominal signal conditions

In the example given, the fade margin is 18.35 dB. Since the transmitter typically contributes very little to the C/N at threshold, M can be approximated by simply calculating the difference between the link C/N under normal signal conditions and the defined minimum link C/N. In the example given, M determined by this method would be approximately $52.6 - 35 = 17.6$ dB.

Based on extensive field testing, the number of hours the path will be below the threshold can be estimated using

$$\text{multipath probability} = U_m = (a)(b)(2.5)(10^{-6})(f)(D^3)(10^{-M/10}) \quad (14.17)$$

where

U_m = the average probability that the excess path attenuation will be greater than the fade margin due to multipath effects
a = a ''terrain factor,'' varying from 4 for very smooth terrain or water to 1 for average terrain to 0.25 for very mountainous terrain

b = a ''temperature/humidity factor,'' varying from 0.5 for very humid areas to 0.25 for normal interior or northern climates to 0.125 for very dry desert or high-altitude areas

f = the operating frequency in GHz

D = the path length in statute miles

M = the fade margin in dB, calculated using Equation (14.16)

When the terrain and temperature/humidity factors are not known, it is common to use 0.25 for the product of a and b for estimation purposes.

14.6.2 Rain Fade

The second factor affecting path availability is rain. Water has a very high dielectric constant as compared with air (about 80 times as high); therefore, when heavy rain cells move through the microwave path, the transmission can be seriously affected. The important parameters are the intensity of the rain, the frequency of rain, the distribution of heavy rain peaks (that is, how big the cells of downpour are), and how fast the cells of intense rain move through the microwave path.

As with multipath, these factors do not lend themselves to theoretical analysis as much as to field experience. Extensive research into rain fade effects at different frequencies and in different geographic areas has been carried out by many researchers. One approach to estimating rain fade uses the methodology outlined next.[6]

First, the available fade margin per kilometer, M', must be calculated by dividing the total fade margin M as determined in Equation (14.15) by the path length in kilometers. Since the path length, D, calculated in Equation (14.3), is in statute miles,

$$M'(\text{dB/km}) = \frac{M\ (\text{dB})}{1.609\,D\ (\text{mi})} \tag{14.18}$$

Second, it is necessary to determine the rainfall rate that will result in an attenuation per kilometer greater than M'. The relationship between signal attenuation and rainfall rate can be approximated[7] using

$$Y_R = kR^{\alpha} \tag{14.19}$$

where

Y_R = the attenuation along the signal path in dB/km

R = the rainfall rate in mm/hr

k and α = ''constants'' that are a function of frequency, polarization, and vertical path angle

Table 14.4 Rainfall Attenuation Constants as a Function of Frequency

Frequency (GHz)	k_h	k_v	α_h	α_v
1	0.0000387	0.0000352	0.912	0.888
2	0.00154	0.000138	0.963	0.923
4	0.00065	0.000591	1.121	1.075
6	0.00175	0.00155	1.308	1.265
7	0.00301	0.00265	1.332	1.312
8	0.00454	0.00395	1.327	1.310
10	0.0101	0.00887	1.276	1.264
12	0.0188	0.0168	1.217	1.200
15	0.0367	0.0335	1.154	1.128
20	0.0751	0.0691	1.099	1.065
25	0.124	0.113	1.061	1.030
30	0.187	0.167	1.021	1.000
35	0.263	0.233	0.979	0.963
40	0.350	0.310	0.939	0.929
45	0.442	0.393	0.903	0.897
50	0.536	0.479	0.873	0.868
60	0.707	0.642	0.826	0.824
70	0.851	0.784	0.793	0.793
80	0.975	0.906	0.769	0.769
90	1.06	0.999	0.753	0.754
100	1.12	1.06	0.743	0.744

k and α, in turn, are determined using the following formulas:

$$k = \frac{k_h + k_v + (k_h - k_v)\cos^2\theta\cos 2T}{2} \tag{14.20}$$

$$\alpha = \frac{k_h\alpha_h + k_v\alpha_v + (k_h\alpha_h - k_v\alpha_v)\cos^2\theta\cos 2T}{2k} \tag{14.21}$$

where k_h, k_v, α_h, and α_v are determined from Table 14.4 and

θ = the path elevation angle in degrees (zero for a horizontal path)

T = the path polarization tilt angle in degrees : 0° for horizontal polarization, 90° for vertical polarization, 45° for circular

For relatively flat paths and horizontal polarization, $k = k_h$ and $\alpha = \alpha_h$; for relatively flat paths with vertical polarization, $k = k_v$ and $\alpha = \alpha_v$.

The maximum allowable rainfall rate can be calculated by solving Equation (14.19) for R and substituting M' for Y_R:

$$R_{\max} = \left[\frac{M'}{k}\right]^{\frac{1}{\alpha}} \tag{14.22}$$

The probability that the rain rate exceeds this value in any given climatic region is given by[8]

$$U_R = 10^{-4}\left[\frac{R_{0.01}}{R_{\max}}\right]^{b} e^{u(R_{0.01}-R_{\max})} = 10^{-4}\left[\frac{e^{(u)R_{0.01}+(b)\ln R_{0.01}}}{e^{(u)R_{\max}+(b)\ln R_{\max}}}\right] \tag{14.23}$$

where u depends on the climate region and is given in Table 14.5 and

$U_R =$ the time-averaged probability that the rainfall rate exceeds $R_{\max}$

$R_{0.01} =$ the rainfall rate, in mm/hr, that is exceeded in any given climate zone for 0.01% of the time (0.86 hr/yr), with an integration time of 1 min. As with u, it is given in Table 14.5.

Table 14.5 Rainfall Variable by Climatic Region

Climate region	u	$R_{0.01}$
A	−0.2391	8
B	−0.0050	12
C	0.05669	15
D	0.10252	19
E	0.03726	22
F	0.04652	28
G	0.07652	30
H	0.04081	32
J	0.1082	35
K	0.0444	42
L	0.03368	60
M	0.04087	63
N	0.02924	95
P	0.02477	145

Note: Figures 14.4, 14.5, and 14.6 delineate worldwide climate regions.[9]

b is determined from

$$b = 8.22(R_{0.01})^{-0.584} \tag{14.24}$$

Once the probability of rainfall exceeding the value needed to attenuate the signal below the specified performance threshold is determined, the annual number of hours of rainfall-caused outage can be determined simply by multiplying U_R by 8,760.

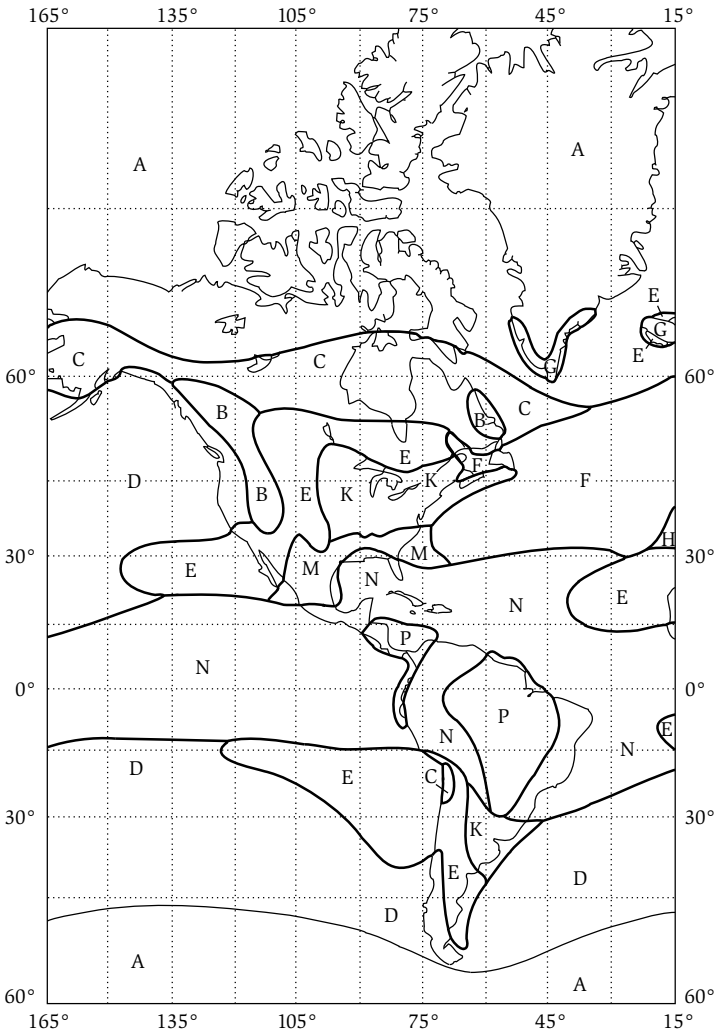


Figure 14.4 Rain zones: the Americas and Greenland.

Finally, once the rainfall and multipath unavailability probabilities are determined, the total unavailability can be expressed as yearly hours of outage using

$$\text{unavailable hours} = 8,760(U_M + U_R) \tag{14.25}$$

Alternatively, the overall path availability can be calculated using

$$\text{path availability} = 1 - U_M - U_R \tag{14.26}$$

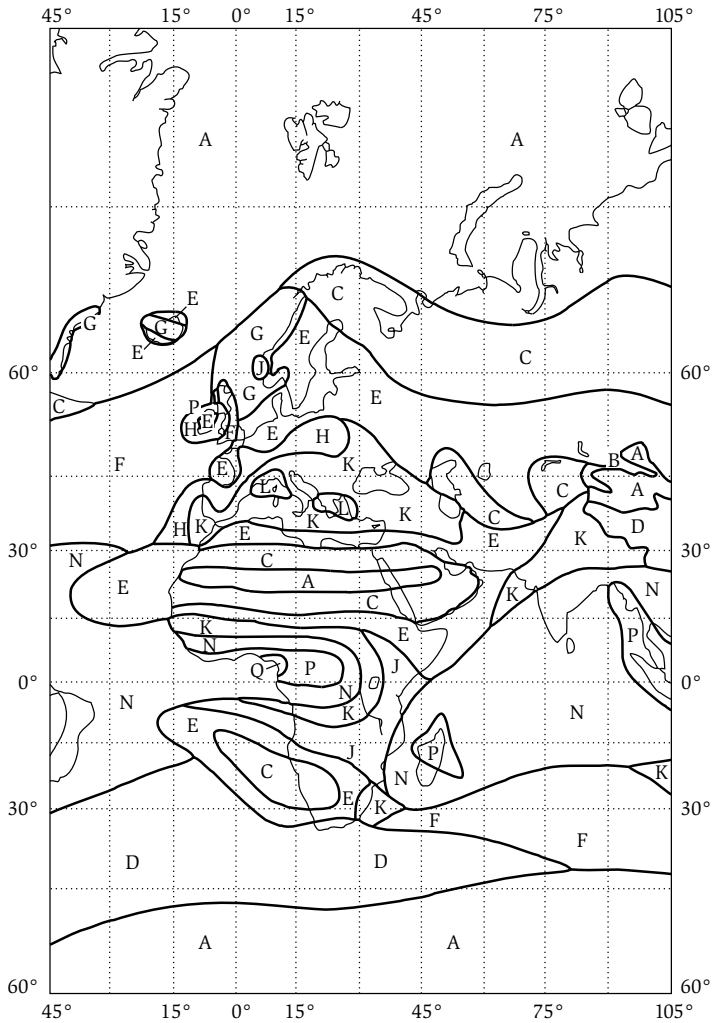


Figure 14.5 Rain zones: Europe and Africa.

Typical link specifications vary from 0.999 availability for noncritical video applications to 0.99999 or even higher for some telephony applications. 0.9999 availability (equivalent to 53 minutes per year average outage time) is a typical goal for a cable television trunking application.

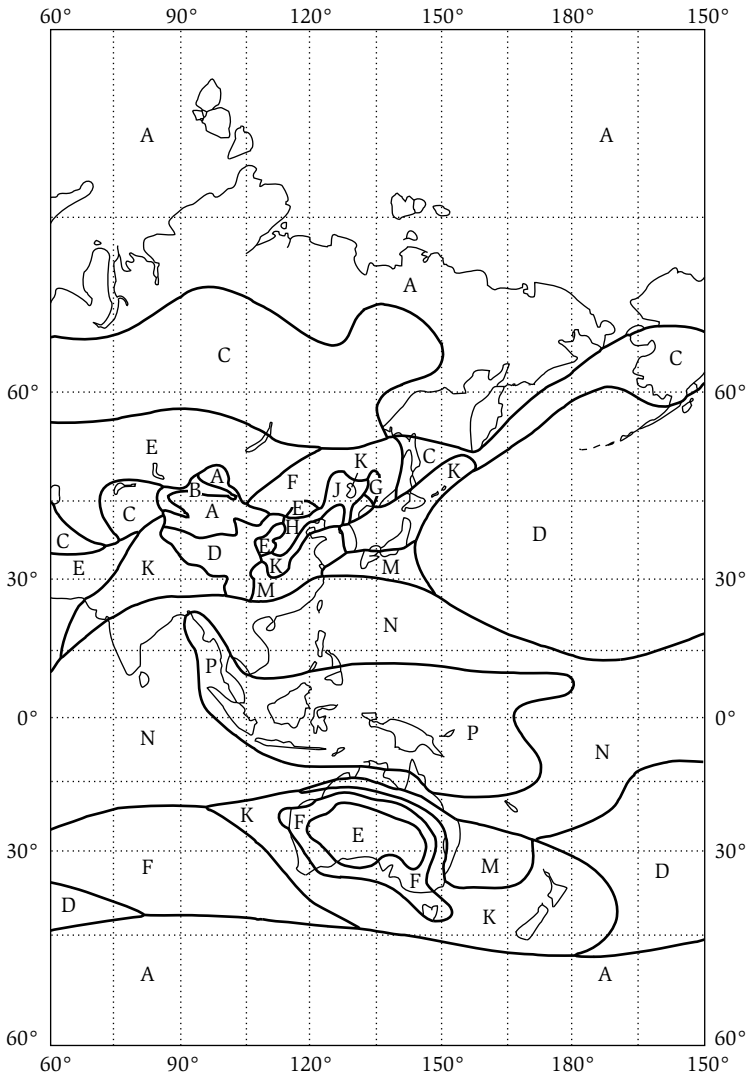


Figure 14.6 Rain zones: Asia and Australia.

14.7 Summary

For most applications, fiber optics is the preferred choice for broadband signal transportation. Microwave links, however, are still a viable and cost-effective way to transport signals where construction costs or physical or regulatory barriers prevent cable construction.

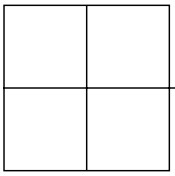
Broadband amplitude modulated microwave links are simple in concept and straightforward to design. The performance of typical AMLs is comparable in noise and distortion to links constructed using directly modulated DFB laser transmitters but suffer from occasional outages due to atmospheric conditions. More important, though fiber-optic links have almost unlimited potential bandwidth, current regulations limit 12-GHz AML links to about 500 MHz of spectrum. The use of higher frequencies, though acceptable for short paths, is limited by the dramatic increase in rainfall and atmospheric attenuation.

The analysis of AML link performance involves a combination of trigonometry and a lot of empirical data on climatic factors, but it is amenable to straightforward computation using simple spreadsheet templates.

Endnotes

1. *Engineering Considerations for Microwave Communications Systems*, G.T.E. Lenkurt, San Carlos, CA, 1975.
2. *Engineering Information*, Rockwell International, Dallas TX, 1979.
3. *The Handbook of Digital Communications*, Microwave Systems News, Vol. 9, No. 11, 1979.
4. *Transmission Applications Guide*, Harris Farinon Division, San Carlos, CA, 1987.
5. C.F.R. 47§17 and C.F.R. 47§78, published by the Office of the Federal Register National Archives and Records Administration and available from the U.S. Government Printing Office.
6. The authors are indebted to Dr. Francisco Bernues of CableAML, Inc., for providing this approach to estimating rain outage probability. Private correspondence, September 1997.
7. From CCIR Report 721–3, p. 229.
8. From CCIR Report 563–3, p. 132.
9. Figures 14.4, 14.5, and 14.6 are extracted from International Telecommunications Union Recommendation ITU-R PN.837–1, *Characteristics of Precipitation for Propagation Modeling*, Figures 1, 2, and 3, and are reprinted with permission of the ITU, which is the copyright holder. The sole responsibility for selecting these extracts lies with the author and can in no way be

attributed to the ITU. The complete volume(s) of the ITU material from which the figures were extracted can be obtained from: International Telecommunications Union, General Secretariat — Sales and Marketing Service, Place des Nations, CH-1211 Geneva 20, Switzerland. They can be contacted on the Internet at: Sales@itu.int.



Chapter 15
End-to-End Performance

15.1 Introduction

The various parts of the broadband distribution system, the trunk, coaxial distribution, and drop, must work together to provide a transmission path whose performance allows delivery of adequate quality signals to the end user. This means that the total system performance requirements must be defined and that the allowable signal degradation must be allocated among signal acquisition, headend processing, distribution system, and terminal equipment.

This chapter will first discuss quality standards, then the question of allocation. Next, it will deal with the calculation of the cascaded broadband network sections. Finally, it will discuss the typical performance of a modern HFC distribution system, including factors that are not normally part of the cascaded network calculations, such as group delay and several classes of undesired signals that affect transmission channels.

15.2 Quality Standards and Requirements

In order to understand how performance standards are related, it is important to first determine the entire chain from signal generation to end user. The components of this chain will vary depending on the service, of course. Some signals may be generated in the headend or even closer to the subscriber, whereas others will be generated thousands of miles away. Figure 15.1 illustrates an example of the transmission elements through which an analog video signal may be sent.

The video source in this example is a television broadcasting station. The source degradation will include some level of white noise, crosstalk, transmitter distortion, and possibly quantizing noise if the video or audio were digitized in any part of the process. Often the (analog NTSC) video S/N is limited by the originating camera and further degraded by a videotape recording and playback process to result in a 55–60-dB input to the

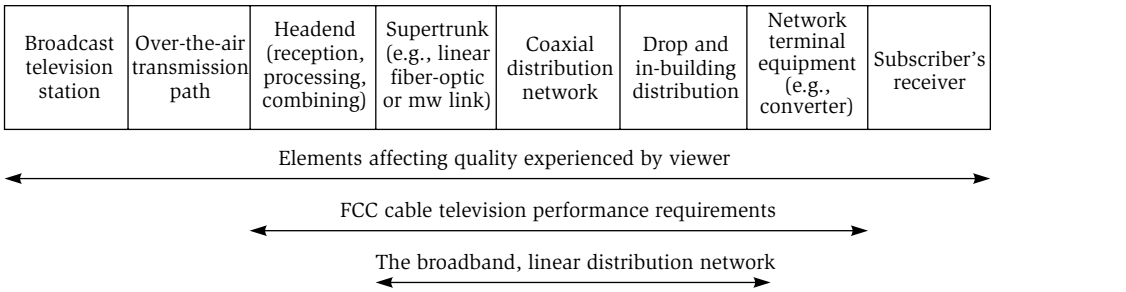

Broadcast television station
Over-the-air transmission path
Headend (reception, processing, combining)
Supertrunk (e.g., linear fiber-optic or mw link)
Coaxial distribution network
Drop and in-building distribution
Network terminal equipment (e.g., converter)
Subscriber's receiver
Elements affecting quality experienced by viewer
FCC cable television performance requirements
The broadband, linear distribution network

Figure 15.1 Various performance measures for a typical broadcast television channel.

transmitter.* The broadcast transmitter diplexing filter (used to combine visual and aural signals) and antenna will affect the frequency response, whereas visual transmitter incidental carrier phase modulation (ICPM) limits attainable stereo audio performance.

The source-to-headend path, if the headend is receiving the signal via the normal transmission path, might include multipath distortion (echoes resulting from signals bouncing off objects to create more than one transmission path), frequency response variations due to receiving antenna characteristics, electrical noise interference, and a host of other factors.

Within the headend complex, the signal may be demodulated, digitized, switched, amplified, combined with other signals, modulated, and/or shifted in frequency. In some cases, signals may be received at a different location than where they are finally combined into the RF spectrum that is transmitted to customers, adding additional layers of processing and transmission. In that case, the term *headend* is intended to include everything from first reception by the network operator to generation of the modulated FDM signal complex for transmission to customers. Each of the processing steps within the headend complex adds a measure of degradation.

Finally the complete frequency spectrum enters the broadband, linear distribution network, where, among others, noise, intermodulation distortion, cross modulation, hum distortion, and group delay are added.

Analog (and possibly digital) video signals may be directly connected to subscribers' receivers, or they may pass through a network termination device, such as an analog or digital set top terminal. If a signal passes through a set top terminal, it may be demodulated, descrambled, and remodulated — each an imperfect process.

---

* In this qualitative discussion, we are not distinguishing between baseband noise (S/N) and noise added to the RF modulated signal (C/N). Both affect the quality of the final product.

Finally, the subscriber's own television receiver will add a measure of degradation. In addition to noise and distortion, it may be inadequately shielded so that off-air signals are mixed with the cable-delivered signals, creating a special form of crosstalk known as *direct pickup* (DPU) interference. DPU results in a variety of visible picture defects in analog signals and may reduce the bit error rate for digital signals.

As shown in Figure 15.1, the viewer sees and hears the composite effect of everything that happens to the signal from source to his or her screen and speakers (or other end application). Chapter 2 describes the results of tests undertaken to determine how overall technical performance in each of several categories is related to subscriber perception of NTSC signal quality.

Obviously, both the chain of network elements and the required quality of signal will be different for services other than analog video. In particular, it appears that various levels of QAM (from QPSK to 256-QAM) will be used to transport digitally modulated downstream signals, whereas lower levels of digital modulation, such as BPSK, QPSK, and possibly 16-QAM, will be used in the upstream direction. The requirements for the various digital signal formats are discussed in detail in Chapters 3, 4, and 5.

Though no end-to-end formal standards have been proposed by the FCC for U.S. television distribution, the Deutsche Bundespost (DBP or German post office — the responsibilities for cable television have since been transferred to Deutsche Telecom) has published standards covering all segments of their entire network from camera to viewer.[1] The standard is very complete with 19 different quality measures covered. By contrast, the FCC regulates few aspects of television broadcast stations (primarily those that assure a lack of interference among stations) but many parameters of cable television systems (including drops if installed and maintained by network operators) and set top terminals (again, only if owned by the network operator). Performance standards for subscriber receivers vary according to how they are marketed. These issues are covered more thoroughly in Part 6. The FCC does not, as of the date of this book, regulate network performance for any signals other than analog video, nor does it regulate the quality of drop materials available for purchase by end users, though voluntary SCTE standards are being developed to cover drop components.

The FCC's standards applying to cable television systems generally specify the total allowable signal degradation from acquisition at the headend to delivery of the final signal to subscribers' receivers (though there are exceptions). This range of coverage is shown by the second horizontal line in Figure 15.1. The quality parameters are applicable (with a couple of minor exceptions) to every analog video signal at every subscriber equipment terminal under all normal operational conditions.

It is important to recognize that the FCC's requirements apply to performance when the network is carrying normally modulated signals. Although these parameters relate directly to subscriber picture quality perception, the

nature of NTSC video is such that some parameters (for example, IM distortion products) are difficult to measure precisely under normal signal conditions because they vary depending on instantaneous levels, carrier phase relationships, and timing of video modulation. Since the total RF power in most cable systems is dominated by analog NTSC video signals, networks are designed and initially tested using unmodulated carriers in place of each expected visual carrier. This yields higher, but more consistent, levels of distortion. Various conversion factors are used to estimate performance under normal signal-loading conditions. In-service systems are tested using normal modulated signal loading, with precision sacrificed to avoid the subscriber service interruption that would be required to perform unmodulated signal testing.

The FCC's operational standards for television broadcasters are contained in the Code of Federal Regulations, Section 47, Part 73, and the cable television regulations are contained in Part 76 and are summarized in Table 15.1. Regulations pertaining to both set top terminals and subscriber receivers are contained in Part 15.[2]

In addition to the FCC's regulations, negotiations between the cable television and consumer electronics industries have resulted in supplemental voluntary standards applying to both cable systems and consumer receivers. Although these were not released as of the writing of this book (and are subject to change), key parameters in draft standard EIA-IS-23 affecting cable system signals delivered to the input terminals of consumers' receivers are as follows:[3]

- The visual level of any individual analog television signal delivered to a receiver should not exceed +20 dBmV, whereas the average of the power levels of all the visual carriers carried should not exceed +15 dBmV.
- When distribution systems use multiple cables, the levels of television signals on the same frequency, but different cables, shall not differ by more than 5 dB.
- The frequency of the visual (luminance) carrier of any analog television signal shall not vary more than ±25 kHz from the nominal values specified in ANSI/EIA-542 (any frequency plan).
- The maximum levels of any non-analog-video signals whose frequencies lie between 0.5 and 30 MHz shall not exceed +42 dBmV, whereas the levels of signals between 54 and 1,000 MHz shall not exceed +20 dBmV (acceptable maximum levels for signals lying between 30 and 54 MHz have not been determined).
- The maximum total average power of nonvideo signals whose frequencies lie between 54 and 1,000 MHz shall not exceed +15 dBmV in any 6-MHz channel.

Table 15.1 FCC rules for analog signals in U.S. cable television systems

Parameter/FCC Rule Paragraph	*Performance Limit*
Visual carrier frequency §76.612	No specific requirement, except that carriers within the specified aeronautical bands (108–137 and 225–400 MHz) must be offset from communications channels with a frequency tolerance of ±5 kHz (see rule for details)
Aural/visual carrier frequency difference §76.605(a)(2)	4.500 MHz ± 5 kHz
Minimum visual signal level §76.605(a)(3)	≥+3 dBmV at the end of a reference 100-foot drop cable connected to any subscriber tap port, and ≥0 dBmV at the input to each subscriber's receiver
Visual signal level stability §76.605(a)(4)	≤8 dB pk–pk variation in visual carrier level over the sum of two 24-hour tests, one taken in July or August and one taken in January or February
Adjacent visual signal level difference §76.605(a)(4)(i)	≤3 dB
Total variation in levels among all visual signals §76.605(a)(4)(ii)	≤10 dB pk–pk for systems with upper frequency limit of 300 MHz or less, plus 1 dB for each additional 100-MHz increment, or fraction thereof (e.g., for a 550-MHz system, the allowed variation is 13 dB)
Maximum visual signal level §76.605(a)(4)(iii)	Below TV overload point (level not specified)
Level of aural signal relative to visual signal §76.605(a)(5)	−10 to −17 dB (−6.5 dB maximum at set top terminal output)
In-channel visual frequency response variation §76.605(a)(6)	≤4 dB pk–pk from 0.5 MHz below the visual carrier to 3.75 MHz above the visual carrier; includes set top terminal response
Visual signal carrier-to-noise ratio §76.605(a)(7)	≥43 dB, measured in a 4-MHz bandwidth
Composite triple beat (CTB) and composite second-order (CSO) IM product levels, relative to visual signal level §76.605(a)(8)	≤−51 dB, time averaged, except ≤−47 dB for products that are frequency coherent with the visual carrier in HRC and IRC systems
Hum modulation §76.605(a)(10)	≤3% pk–pk modulation of the visual carrier (≤−1.5% modulation by conventional definition)
Relative chrominance/luminance transmission delay (chroma delay) §76.605(a)(11)(i)	≤ ± 170 ns, not including the contribution of any set top terminal
Differential gain §76.605(a)(11)(ii)	≤ ± 20%, not including the contribution of any set top terminal
Differential phase §76.605(a)(11)(iii)	≤ ± 10 degrees, not including the contribution of any set top terminal

Regardless of their official status, these additional performance guidelines represent good engineering practice in assuring that cable television signals are compatible with most consumer electronics receivers.

Not included in either U.S. government regulation or interindustry agreements are specifications covering several aspects of the sound accompanying NTSC signals (such as loudness level or stereo separation) though these are included in the DBP model. An issue with both over-the-air broadcast and cable-generated video programming has been the perceived loudness variations among channels and within channels (for example, between the main program and commercials). Despite several attempts by both governmental and industry groups, it has never been successfully resolved.

Also not specified in any document are limitations on either microreflections or external signal ingress. Microreflections were discussed in Chapter 11 and result in visible ghosts in analog NTSC pictures if the amplitude and delay are sufficient. As will be seen later in this chapter, they also contribute to both group delay and frequency response variations. External signal ingress can occur in either the distribution network or within subscribers' terminal equipment. Signal pickup in subscribers' receivers, discussed in Chapter 24, can affect the network if ingressing signals are transmitted out the antenna terminals. It should be noted that, though the FCC does have strict standards covering signal egress from cable systems, where strong external signals are present, ingress will usually be visible to subscribers before the FCC *egress* limit is exceeded though the two effects both arise from inadequate shielding and are obviously related.

Several voluntary industry standards exist for digitally modulated signals. The interindustry specification EIA-105, Decoder Interface, for example, includes several tuner performance requirements that are intended to ensure successful reception of high-order VSB or QAM signals; the SCTE's Digital Video Subcommittee and Advanced Television Systems Committee are developing digital video transmission standards, and the data-over-cable service interface specification (DOCSIS) covers cable modems. Chapter 4 covers digital modulation in detail, along with the required transmission quality.

15.3 Performance Allocations Among Sections of the Cable System

Among the listed parameters, clearly some are determined exclusively (or primarily) in one element of the transmission chain, whereas others (such as carrier to noise) are cumulative as the signal travels through the sections. Specifically:

- The visual carrier frequency is initially determined at the point of RF modulation but may be modified by subsequent frequency conversions. In most set top terminals, the video signal is demodulated and remodulated so that the frequency of the signal delivered to the subscriber's

equipment is not related to that on the network. In that case, the FCC's frequency standards apply to both the headend output and the output of each terminal.

- The difference between visual and aural carrier frequencies is also determined at the point of modulation and is unaffected by subsequent frequency conversions that modify the entire channel's spectrum (or the entire RF spectrum, for instance, in an AML transmitter). Similarly, the deviation of the aural carrier (and therefore the "loudness" of the detected sound) is determined at the last point of modulation, be that the over-the-air transmitter, headend, or set top terminal. Here again, the FCC's rules apply at the last point of modulation.
- The levels of signals delivered to subscribers' homes are primarily a function of the broadband distribution system but are also affected by changes in the headend. Where set top terminals are used, however, the signal levels reaching subscribers' receivers may have no relationship to system operating levels. The FCC rules on levels delivered to subscriber equipment apply both before and after any operator-supplied terminal equipment. In addition to the Part 76 rules covering cable television, specifications in Part 15 further limit the allowable output RF power range of terminals.
- In-channel frequency response variation, relative visual/aural carrier level difference, chroma delay, differential gain, and differential phase in a properly operating cable system are primarily determined in the headend and only slightly modified as a result of variations in the response of the broadband network. As with visual levels, however, set top terminals significantly modify all these parameters.
- Intermodulation distortions (CTB, CSO, XMOD, and CIN) primarily occur in the broadband distribution network but can be further degraded by set top terminals.
- Hum modulation can occur at any place in the network though, in a properly maintained system, it will primarily occur because of the side effects of ac powering being carried over the same coaxial system as RF signals — a combination of power pack ripple and parametric modulation of magnetic components.
- Noise is added at all stages in the process though not in equal amounts.

In summary, the primary impediments that occur in multiple sections of cable systems are distortion (occurring in both broadband distribution network and terminal equipment) and noise (with contribution from headend, broadband distribution network, and terminal equipment). With this understanding, we can discuss allocations of the allowable degree of cable system signal degradation among the various major network sections and within the subparts of the broadband distribution section.

15.4 Noise and Distortion Allocations in Cable Systems

15.4.1 Carrier-to-Noise Ratio

Applying Equation (11.2) to the elements of a cable system, it can be seen that the total C/N measured from signal acquisition to the input of subscribers' receivers is

$$C/N_{ttl} = -10\log\left[10^{-\left(\frac{C/N_h}{10}\right)} + 10^{-\left(\frac{C/N_s}{10}\right)} + 10^{-\left(\frac{C/N_d}{10}\right)} + 10^{-\left(\frac{C/N_t}{10}\right)}\right] \qquad (15.1)$$

where

C/N_{ttl} = the system C/N considering only noise contributions from the headend input to the input of subscribers' receivers

C/N_h = the C/N of the headend, considered alone

C/N_s = the C/N of the supertrunk (generally fiber-optic or microwave link), considered alone

C/N_d = the C/N of the coaxial distribution, considered alone

C/N_t = the C/N of the terminal equipment (for example, set top terminal), considered alone

Although this formula appears somewhat complex, it reflects converting each section's relative noise level to a scalar value, adding the normalized noise levels, and reconverting to familiar logarithmic (dB) terms. The initial 10 and the 10 used to divide each of the original C/N values reflect the fact that the thermal noise generated in each section is noncorrelated and so adds on a power rather than voltage basis. A shorthand way of stating this relationship is to state that the "cascade factor" is "10 log."

One eminent cable engineer has suggested the use of a shorthand notation to express such equations, which we have slightly modified to make it more universal.[4] Using this notation, Equation (15.1) can more simply be written as

$$C/N_{ttl} = C/N_h \oplus C/N_s \oplus C/N_d \oplus C/N_t : 10\log \qquad (15.2)$$

where the $\oplus$ sign indicates the conversion to scalar values, addition, and reconversion to dB form, and the term after the colon indicates the cascade factor, which may vary from 10 log (totally uncorrelated power addition) to 20 log (voltage addition). We will use this notation throughout this section.

Typical set top terminals have noise figure specifications ranging from 10 to 13 dB. Since cable operators are only required to deliver analog television signals

at levels of 0 dBmV (and typically deliver many channels at close to that level), we can use Equation (10.19) to determine that, with the noisiest terminals and minimum signal levels:

$$C/N_t = 0\,\text{dBmV} + 59.2\,\text{dB} - 13\,\text{dB} \simeq 46\,\text{dB} \tag{15.3}$$

At the other end of the chain, individual modulators and signal processors will generally exhibit 60-dB C/N or better. That is degraded, however, by the out-of-band noise from other modulators when many channels are passively combined. The various summing and isolation amplifiers add additional noise as well (see Chapter 8 for a discussion of required headend processing) so that the typical C/N of individual channels as measured at the input to the broadband distribution network is of the order of 57 dB.

Assuming that the goal is to just meet the FCC's required C/N_{ttl} of 43 dB, we can find the minimum composite C/N of the supertrunk and coaxial distribution together using

$$\begin{aligned} C/N_s \oplus C/N_d = C/N_{ttl} \ominus C/N_h \ominus C/N_t : 10\log \\ = 43 \ominus 57 \ominus 46 = 46.4\,\text{dB} \end{aligned} \tag{15.4}$$

where the $\ominus$ sign indicates that the scalar quantities are subtracted rather than added.

What this equation shows is that our worst-case terminal has effectively used up half the total noise budget of the system. Though most modern terminals have lower noise figures, set top terminals still contribute significantly to total noise.

Typical design specifications for the supertrunk plus coaxial distribution portion of the plant call for 48–49 dB C/N. The margin of about 2 dB provides an allowance for such operating variations as errors in setting operating levels at the headend, aging of components, and imperfect frequency response through chains of amplifiers. The variation in frequency response is known as peak-to-valley (P/V) and is a measure of the peak-to-peak variation, in decibels, from the ideal. Though fiber-optic links are typically flat within a few tenths of a decibel, the usual allowance for coaxial distribution networks is $N/10 + 2$, where N is the number of cascaded amplifiers.

The allocation of noise budget between supertrunk and coaxial distribution depends on both architecture and technology. Typical downstream fiber-optic links driven by directly modulated DFB transmitters will exhibit C/N values of 51 to 54 dB, as will broadband AML microwave links. In a simple architecture, a 49-dB supertrunk/coaxial C/N requirement might be met by cascading fiber-optic and coaxial sections, each independently providing 52-dB C/N. More complex architectures might require cascaded supertrunk links, requiring better performance in each section.

15.4.2 CTB and CSO

Since headend levels can be carefully optimized and premium amplifiers utilized, intermodulation distortion within the headend is minimized, and most distortion occurs in the supertrunk, coaxial distribution, and set top terminal. As discussed in Chapter 11, however, the calculation is not straightforward. Continuing the example of a simple HFC network, the distortion mechanisms in the directly modulated DFB laser, the balanced RF amplifiers, and the input stages of the set top terminal have some characteristics in common (all exhibit some form of limiting) and others that are different (the small-signal linearity variations and nature of the limiting when it does occur).

Using the same notation as in Figure 15.1, we can write a general equation for the composite distortion:

$$C/dist_{ttl} = C/dist_s \oplus C/dist_d \oplus C/dist_t{:}nlog \tag{15.5}$$

where *dist* could indicate CTB or CSO and n is the cascade factor, which can vary from 10 to 20. When the dominant distortion mechanism is similar in the cascaded elements, for instance, symmetrical compression in cascaded push–pull amplifiers running at the same output levels, the cascade factor will be close to 20, which was the number traditionally used for both CSO and CTB in large all-coaxial networks. Research in the performance of dissimilar cascaded sections, however, has shown that distortions often combine at less than a voltage addition rate (see Endnote 2, Chapter 11). In occasional cases, in fact, the distortion may build up at a less than 10 log rate when distortions partially cancel.

As discussed in Chapter 11, the broadband distribution network is designed and (often) tested using unmodulated carriers. This results in CTB and XMOD product design levels that are worse than average operational values by about 12 dB, whereas CSO products will be worse than operational values by about 6 dB, though the difference may not be that great for very short cascades or where the modulation on multiple channels is synchronized.

Although there are no standards in this area, common industry practice has been to design for about 53-dB C/CTB and C/CSO in the broadband network (supertrunk plus coaxial distribution) under test (unmodulated carrier-loading) conditions. Even when combined with typical set top terminal distortion levels (57–65-dB C/CTB and 60-dB C/CSO), the total distortion under operational conditions is well within FCC requirements, even with the most conservative cascade factor assumptions, as shown in these equations:

$$\begin{aligned} C/CTB_{ttl} &= (53 \oplus 57{:}nlog) - 12\,\text{dB} \\ &= 60.8\,\text{dB for } n = 20 \end{aligned} \tag{15.6}$$

$$\begin{aligned} C/CSO_{ttl} &= (53 \oplus 60{:}nlog) - 6\,\text{dB} \\ &= 55.8\,\text{dB for } n = 20 \end{aligned} \tag{15.7}$$

Typical specifications for directly modulated DFB optical transmitters include a C/CTB of 65 dB and C/CSO of 62 dB. Externally modulated transmitters are generally better, perhaps 70-dB C/CTB and 65-dB C/CSO, whereas broadband microwave links may generate slightly worse distortion.

The remaining distortion "budget" can be allocated to the coaxial distribution system or to additional supertrunk links, if required by the architecture. Fortunately, though fiber-optic links have higher CSO than CTB, coaxial amplifiers have higher CTB so that the cascade of the two technologies is somewhat complementary with regard to distortion.

Although we could do a similar calculation for XMOD, the results would generally track CTB since both are third-order effects. In practice, CTB testing has proved to produce more consistent results and so is generally preferred to XMOD evaluation. It should be noted that the FCC no longer specifies XMOD performance.

15.5 Typical Network Transmission Quality Under Operational Conditions

Having discussed how various signal imperfections arise in the network and how they are designed and tested, we will now consider, as an example, typical HFC network performance under the following conditions:

- 54- to 750-MHz total downstream frequency range, loaded with 78 analog NTSC channels extending up to 550 MHz, plus 200 MHz of digitally modulated signals extending to 750 MHz
- A single fiber link extending from the headend to each node and feeding a coaxial distribution network that passes 500–2,000 homes
- A nominally flat frequency response from the headend output through the fiber link and to the input of the coaxial distribution system; an interstage circuit to set an optimum gain/frequency slope within the first amplifier; and a cascade of amplifiers, interconnecting cable, and passive devices adjusted so that the frequency response from the output of any amplifier to the output of the subsequent amplifier is flat (see Figure 10.19 for a graphical illustration)
- Equal levels among the analog video channels leaving the headend and the levels of digitally modulated carriers set such that the total average power per 6-MHz channel is 10 dB below the levels of the analog video visual carriers during synchronizing peaks.

These parameters are typical of specifications, maximum signal-loading capabilities, and operational parameters of many cable systems built in the mid-1990s. Most coaxial and fiber-optic equipment designed for this application is specified

under both the preceding carriage conditions and, alternately, under a load of 110 NTSC video signals.

As discussed in Chapter 13, it is not always the case that digitally modulated signals and analog video signals share the same optical fibers on transmitters. Nevertheless, the methodology discussed in this chapter can easily be applied to more complex systems.

This section should provide guidance on typical end-to-end transmission channel conditions to those designing products whose signals will be transmitted through cable systems, as well as general guidance on how operating performance is related to design parameters. An important caveat, however, is that although the cable television industry has now deployed HFC architectures for most of its networks, many older all-coaxial networks are still in use, and new networks are being constructed with more complex architectures both with and without microwave supertrunks. These differences must be taken into account. As an example, microwave links are subject to such factors as fading, whereas all-coaxial networks will exhibit greater level instability and in-band response variations. Either technology, however, will outperform directly modulated DFB laser transmitters with regard to clipping distortion.

Each of the channel impairments summarized here will be expressed relative to two kinds of signals:

- Analog video signals that originate within or are received at the headend and that pass through the network and a set top terminal before delivery to customers' receivers
- Digital signals that are inserted at the last combining point in the headend, operate at 10-dB depressed levels (as defined earlier), and are delivered directly to terminating equipment

Given these two classes of signals, it is simple to estimate the effect on other signals of different characteristics.

15.5.1 Carrier-to-Noise Ratio

The FCC's video C/N specification is one of the most difficult to meet for the most distant (in terms of equipment cascade) cable system subscribers. For those subscribers who also use set top terminals, it is likely that the required 43-dB C/N is barely met on some channels. For customers who are located close to a fiber node and who directly connect their receivers to cable outlets, the C/N may be more than 50 dB (limited by the noise contributions of the headend, fiber link, and one coaxial amplifier).

A digitally modulated signal that is inserted at the input to the downstream laser transmitter (and thus avoids the noise contributions of headend amplifiers), which is operated at a level 10 dB below equivalent analog video channels and is

directly connected to termination equipment, is affected by different factors. The C/N at the most distant subscriber tap for such signals is essentially that of the supertrunk and coaxial distribution network, corrected for the signal level difference between video and data signals, corrected for the noise susceptibility bandwidth of the digital receiver relative to 4 MHz, and with an added allowance for P/V, errors in setting levels, and aging of components. As with analog video, the total variation due to these factors may be of the order of 2 dB. Thus, we can write a generic equation for likely worst-case C/N for a non-analog-video signal:

$$C/N_{\text{digital}} = (C/N_s \oplus C/N_d : 10\log) - S - 10\log\left(\frac{B}{4}\right) - \text{margin} \qquad (15.8)$$

where

$C/N_s =$ the analog video C/N in the supertrunk
$C/N_d =$ the analog video C/N in the coaxial distribution
$S =$ the suppression of digital signal levels relative to analog video (in dB)
$B =$ the noise susceptibility bandwidth of the digital receiver (in MHz)
margin = the expected variation from design performance due to aging, P/V, and operational tolerances (in dB)

As an example, if $C/N_s \oplus C/N_d$ is 48 dB, the signals are run 10 dB below video, the bandwidth is 6 MHz, and we allow a 3-dB margin, the worst-case C/N_{digital} would be 33.2 dB.

15.5.2 Intermodulation Products from Discrete Carriers

As discussed earlier, typical broadband networks (supertrunk plus distribution) are designed to provide C/CTB and C/CSO levels of at least 53 dB under unmodulated signal test conditions. Combined with hypothetical set top terminal distortion levels and adjusted for the difference between unmodulated carriers and normal television signals, worst-case C/CTB will be about 61 dB, whereas C/CSO will be about 56 dB.

As with C/N, a further allowance must be made for imperfect frequency response (P/V) and for operational variances and aging of components. Errors in setting signal levels will cause larger differences for CTB since the distortion product amplitudes change by 3 dB for every 1-dB change in desired levels, whereas CSO products change by only 2 dB. Furthermore, the levels of each video signal and the total peak RF power carried will change over time as average picture brightness changes and as the various video-modulating waveforms drift into and out of time synchronization. The result is that the observed product clusters affecting any one channel will vary by several decibels over

time (the FCC allows time-averaged tests in evaluating operational systems). If we allow 5 dB for the total of these effects on CTB and 3 dB for the CSO effects, we can predict that the worst-case C/CTB on any channel at any time will be about 55 dB, whereas the worst-case C/CSO will be about 53 dB.

The improvement in performance experienced by subscribers who are near optical nodes is not as great as for C/N since the largest contributor to CSO is the optical transmitter, whereas the largest contributor to CTB is generally the last amplifier (or two) that drive each subscriber tap string.

In the case of digitally modulated signals sharing the cable network with analog television signals, some intermodulation products will fall into the digital spectrum. If the same channelization is used, then the products will appear in the same relative positions within each channel. That is, for ANSI/EIA-542 Standard channels, they will occur at harmonics of 6 MHz and 1.25 and 2.5 MHz above those harmonics, with the first and third of those being CSO products and the middle one being the CTB product. Scaling from the analysis just done for video signals, adjusting for the 10 dB of level difference, and eliminating the contribution from the set top terminal (approximately 69-dB C/CTB and 66-dB C/CSO with modulated signals), we would expect that the worst-case C/CTB might be 47 dB, whereas the worst-case C/CSO might be 45 dB.

Given the distribution of CSO products with frequency (Figure 10.8), it would be expected, further, that the only significant CSO products would be those occupying the highest frequency in each set (2.5 MHz above the harmonics of 6 MHz) when digital signals are placed at the top end of the spectrum. ANSI/EIA-542 Standard chennels are further protected against lower-side CSO products because they occur only at the boundary between channels.

Though other ANSI/EIA-542 channelization alternatives are declining in popularity, it should be mentioned that with the HRC system, all the second- and third-order products will fall exactly on harmonics of the reference oscillator (generally 6.00003 MHz), as will the visual carrier, if present. Since the products are frequency coherent, they do not cause visible beats in analog channels and so are less visible. With the IRC system, the products have approximately the same frequencies as with Standard channelization, except that each product group is frequency coherent. Because the second-order products are not coherent with the visual carrier, second-order beats may become visible in analog channels carried in IRC systems before third-order beats in some cases.

15.5.3 Composite Intermodulation Noise (CIN)

As discussed in Chapter 10, intermodulation products among digital signals and between digital signals and analog visual carriers add “bands” of noiselike products, which add to thermal noise to raise the effective noise floor of the system. The frequency distribution of these products is a function of the distribution and relative power levels of both the analog and digital signals. Fortunately, when the average power density of digital signals is suppressed by 5 dB

or more relative to visual carriers, the effective C/N degradation for the visual signals is typically less than 1 dB.

15.5.4 Laser Transmitter Clipping Distortion

As discussed in Chapter 12, downstream laser clipping distortion is a statistically rare event in a properly adjusted, directly modulated DFB transmitter. When external modulation methods are used, clipping is even rarer since other nonlinearities limit average modulation levels before clipping probability becomes significant. Thus, it can reasonably be expected that laser clipping will have no measurable, visible, or audible effect on analog NTSC channels under normal operating conditions. Should the optical modulation levels be adjusted too high, however, horizontal black streaks will appear in pictures.

Studies of clipping-induced errors in 64-QAM data signals that share DFB optical transmitters with a full spectrum (54–550 MHz) of analog video signals have shown that, with proper transmitter adjustment, bit error rates due to clipping are less than 10^{-9} even without forward error correction. If, however, the modulation is set too high, the BER can increase to the order of 10^{-7} or even lower before affecting subjective picture quality. Once the threshold of clipping is approached, the decline in performance with increasing modulation is very steep.

15.5.5 Hum Modulation

Hum modulation occurs in both powered equipment (due to imperfect power supply filtering) and as a result of parametric modulation of magnetic components in both amplifiers and passive devices. As discussed in Chapter 10, even though the FCC allows 3% pk–pk modulation levels, systems seldom exceed half that in practice. Modulation due to power supply ripple should logically be more or less independent of RF frequency. To the extent that parametric modulation of magnetic components is a factor, however, its effects may vary with frequency. The primary mechanism is that magnetic cores may partially saturate at peaks of the ac current. This will affect their impedance at RF, and so create a change in return and transmission loss that will vary at a 60-Hz rate. Since the impedance change may vary with frequency, as may the effect of the impedance change, there is no way of predicting how the modulation percentage will vary across the spectrum. At a given frequency, however, the percentage modulation of a video signal should be the same as that of a digitally modulated signal though the latter may be able to tolerate a higher modulation percentage.

15.5.6 Microreflections

Microreflections can occur anywhere within the broadband coaxial network (or, for that matter, in the optical link) but are most severe where components

are mounted within a few hundred feet of each other, namely, in the tapped coaxial feeder lines and drop structure. Even though component quality (and thus impedance match, isolation, and directivity) is higher in the "hard cable" plant (that is, the coaxial portion on the network side of subscriber taps), that is partially offset by the lower per-foot cable losses and greater reflection delays.

The mechanisms by which microreflections occur were analyzed in Chapter 10. Drops were analyzed in Chapter 11. In summary, occasional microreflections should be expected whose magnitude can vary from as large as −10 dBc at zero delay to a maximum amplitude that will decline in accordance with round-trip cable losses at the frequency in question.

Such large reflections, however, can only occur in the trunk and distribution network in the case of a major cable or component fault, poor-quality passive components, or a terminating tap, all of whose subscriber ports are unterminated. If any tap port is terminated, even by connection to a drop cable, the reflections will be significantly smaller. In the drop structure, reflections from receivers and unterminated outlets are partially compensated for by the larger attenuation of the interconnecting cables. As a practical rule, most delayed signals should be at least 20 dB below direct signals, with delays ranging from 0 to about 500 ns. In addition to reflections in tapped lines and within the drop structure, mismatches occurring in long, untapped cable spans (trunk or express feeder lines) can result in smaller reflections (≤ -40 dBc) with delays as long as 5 μs.

Microreflections can cause various picture degradations in analog transmitted signals, ranging from subjective softening or overemphasis of vertical lines (depending on the relative polarity of the reflected signal) when delay times are short to visible ghosts when the delay time is sufficient that the reflected picture is distinguishable from the direct picture. Customer perception of microreflections' effect on picture quality is discussed in Chapter 2.

With data signals, the reflected signal may be interpreted in a number of ways. In the time domain, the effect is to create an amplitude uncertainty in detected amplitude and phase as the delayed and direct signals add at the detector. In the frequency domain, these same effects can be interpreted as group delay and in-band response variations.

15.5.7 Group Delay and In-Band Response Variations

To review, the return loss of a component is the ratio between the signal impinging on a port (with all other ports properly terminated) and the signal reflected from that port:

$$R\,(\text{dB}) = L_i\,(\text{dBmV}) - L_r\,(\text{dBmV}) = 10\log\left[\frac{P_i}{P_r}\right] \tag{15.9}$$

where

R = the return loss in dB

L_i and L_r = input and reflected power levels in dBmV (or any other consistent logarithmic units)

P_i and P_r = input and reflected power levels in mW (or any other scalar quantities)

Another way of quantifying the mismatch recognizes that the voltage due to the reflected wave will, at some points along the transmission line, add to the incident wave, whereas at others it will subtract from it. The ratio of maximum to minimum combined voltage is known as *voltage standing wave ratio* (VSWR). Note that although at any point the voltage will vary in a sinusoidal fashion at the frequency of the signal, the *ratio* of voltages along the line will not change, thus a "standing" wave ratio. The relationship between VSWR and return loss is

$$VSWR = \frac{(1 + 10^{-R/20})}{(1 - 10^{-R/20})} \tag{15.10}$$

Figure 15.2 illustrates the standing waves resulting when a transmission line is terminated in a pure resistance equal to half its characteristic impedance.

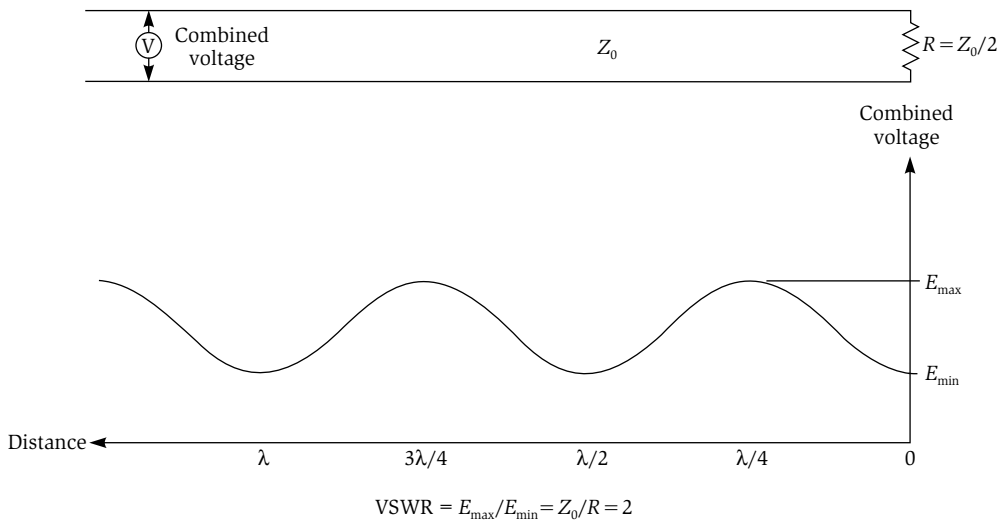


Figure 15.2 Standing wave pattern.

When a line is terminated in a pure resistance of any value, the VSWR is Z_0/R or R/Z_0, whichever is greater.

As can be seen, for purely resistive terminations, the standing wave pattern has a maximum value at distances that are odd multiples of one-quarter wavelength from the termination and a minimum value at distances that are even multiples. The reverse would be true if R were greater than Z_0. If the termination is complex (not a pure resistance), then the distance to the first minimum or maximum will be other than a quarter wavelength. Regardless of the termination, however, the maxima and minima will be exactly one-quarter wavelength apart at the frequency of the signal.

The phase shift, ϕ, through a length of cable, in radians, is $2\pi L/\lambda$, where L is the length and λ is the wavelength in the cable. Substituting for λ from Equation (10.9):

$$\phi = \frac{2\pi L f}{984 V_P} \tag{15.11}$$

where

ϕ = the phase shift in radians
L = the distance, in feet, over which the phase shift is measured
f = the frequency in MHz
V_P = the relative velocity of propagation in the cable

We would like to know what frequency shift, Δf, would be required to cause a relative phase difference of 2π radians through a fixed length L of cable. If we solve Equation (15.11) for f and let $\phi = 2\pi$, we get

$$\Delta f = \frac{984 V_P}{L} \tag{15.12}$$

If a direct signal and one delayed by being transmitted through a cable of length L are mixed, as through a reflection, then Δf is known as the ripple frequency.

Reflections Due to Mismatches in Trunk and Distribution Plant

To see how this affects amplitude response and group delay, let us examine a portion of a typical tapped distribution line (Figure 15.3). Signals exit the source, which might be an upstream tap or amplifier. They travel through cable 1 and enter the tap. Signals exiting the tap through-port travel through cable 2 to the next downstream component, labeled load, which might be another tap or the input to a following amplifier.

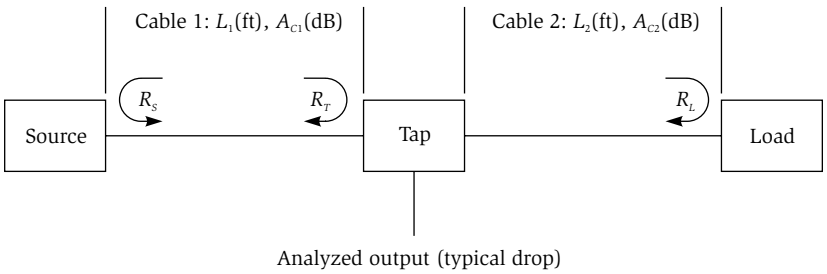


Figure 15.3 Sample of tapped line segment.

The following parameters will be used in the calculations:

Length of cable 1 in feet $= L_1$
Length of cable 2 in feet $= L_2$
Source return loss in dB $= R_s$
Load return loss in dB $= R_L$
Tap return loss in dB $= R_T$
Attenuation of cable 1 in dB $= A_{C1}$
Attenuation of cable 2 in dB $= A_{C2}$
Tap directivity in dB $= D$
Frequency in MHz $= f$

Reflected power will be coupled into the drop by two primary mechanisms:

- A portion of the signals that are reflected from the tap input will be re-reflected from the source and combined with the direct signals delayed by the transit time through $2L_1$ feet of cable. The difference in level between the desired and reflected signals, R_1 (in dB), will be $2A_{C1} + R_S + R_T$.
- A portion of the signals that are reflected from the load will be coupled into the drop due to the finite directivity of the tap. These signals will be delayed relative to the desired signals by the transit time through $2L_2$ feet of cable. The difference, R_2, in level between the desired and delayed signals in this case will be $2L_{C2} + R_L + D$.

In either case, the ratio of the scalar magnitudes of the desired, E_F, and delayed, E_D, signal voltages appearing at the drop port will be $10^{R/20}$.

The presence of the reflected/delayed signal affects the desired signal in three ways. First, the delay, if longer than the symbol period of a digital transmission, will reduce the error threshold by interfering with the following symbol. In the case of an analog television signal, the visible effect is a horizontally offset image whose visual appearance depends on the amplitude and delay time, as covered previously.

Second, the frequency-sensitive nature of the VSWR pattern between the direct and reflected signals will affect the apparent amplitude versus frequency response of the system. Using Equation (15.10), the peak-to-peak change in amplitude response can be calculated to be

$$\text{ripple (dB pk–pk)} = 20\log\left[\frac{(1+10^{-R/20})}{(1-10^{-R/20})}\right] \tag{15.13}$$

This ripple will occur at the incremental frequency rate calculated in Equation (15.12), where L is either $2L_1$ or $2L_2$ depending on the dominant coupling mechanism.

Finally, the vector combination of the desired and delayed signals incrementally affects the phase shift through the system, and thus group delay. Figure 15.4 shows the relationship between the magnitudes of the desired and reflected signals, the angle between them, and the angle and magnitude of the resultant signal. In general:

$$\begin{aligned}\theta &= \arctan\left[\frac{E_D\sin\phi}{E_F + E_D\cos\phi}\right] \\ &= \arctan\left[\frac{10^{-R/20}\sin\left(\dfrac{2\pi Lf}{984V_P}\right)}{1+10^{-R/20}\cos\left(\dfrac{2\pi Lf}{984V_P}\right)}\right]\end{aligned} \tag{15.14}$$

We can restate the definition of group delay (Equation (10.32)) in similar units:

$$\text{group delay (ns)} \equiv \frac{1{,}000}{2\pi}\frac{d\theta}{df} \tag{15.15}$$

where

θ = the incremental phase shift in radians

$d\theta/df$ = the rate of change of phase shift through the device as a function of frequency in units of radians per MHz

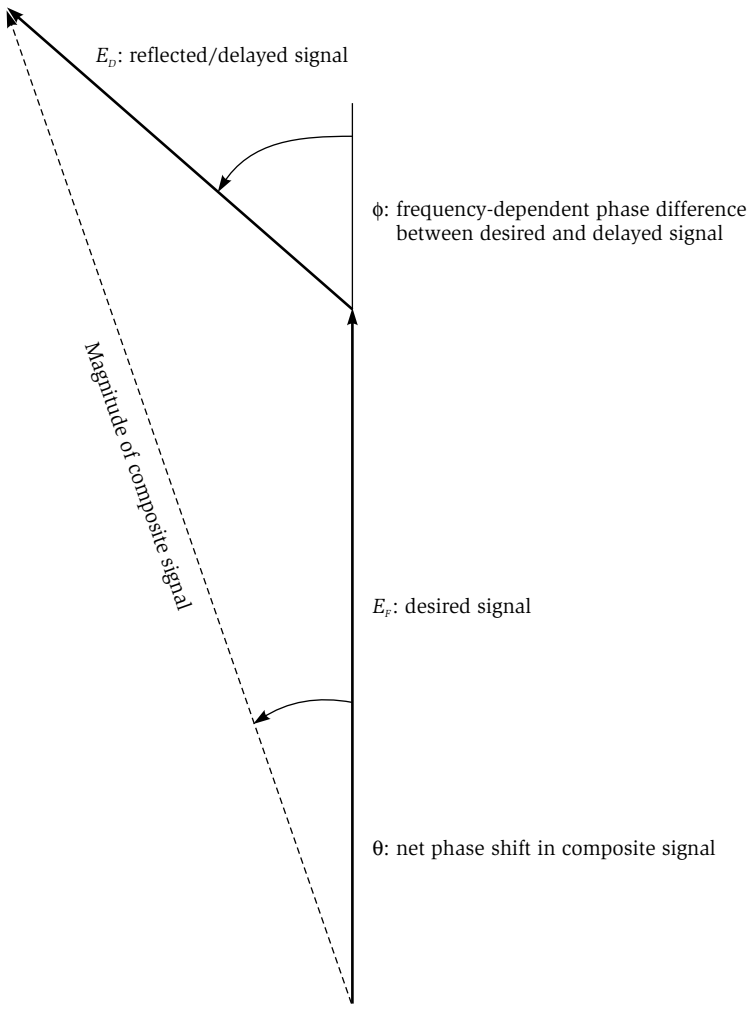

E_D: reflected/delayed signal
ϕ: frequency-dependent phase difference between desired and delayed signal
Magnitude of composite signal
E_F: desired signal
θ: net phase shift in composite signal

Figure 15.4 Vector addition of desired and reflected/delayed signals.

This allows us to solve for the peak-to-peak delay variation due to the presence of the reflected signal:

$$\text{group delay variation (ns pk}-\text{pk)} = \left[\frac{2.032L}{V_P}\right]\left[\frac{10^{R/20}}{10^{R/10}-1}\right] \qquad (15.16)$$

where

L = the incremental cable length through which the reflected signal is transmitted in feet

V_P = the relative propagation velocity (dimensionless)

R = the ratio of the desired to reflected signal strength as measured at the drop port and expressed in dB

A practical example may be useful in quantizing the effects of reflected signals in typical coaxial distribution lines. Typical amplifiers and passives have return loss specifications of about 16 dB when properly terminated, with the worst performance near the ends of the passband, as can be expected. The specified directivity of taps varies as a function of both frequency and tap value. In general, the worst values are also at the extremes of the passband (especially at the low end of the upstream frequency range) and at high tap values, where the absolute coupling between ports limits the isolation. For most values, the directivity in the downstream spectrum is 10–15 dB, whereas the return band directivity is 6–9 dB. Finally, most distribution lines use cable whose loss is about 1.6 dB/100 feet at 750 MHz and varies, as we have seen, as the square root of frequency. The relative velocity of propagation for these cables is about 0.90.

Given these parameters, assume that an amplifier and two taps are spaced at 100-foot increments along a cable. At 750 MHz, the following will be the situation under proper operating conditions:

- The signal reflected from the first tap and re-reflected from the amplifier will be attenuated by $16 + 16 + 2(1.6) = 35.6$ dB and delayed by $200/[(0.984)(0.9)] = 226$ ns. The signal transmitted through the first tap, reflected from the second tap and coupled into the drop due to a directivity (which we will assume to be 12 dB), will be attenuated by 31.6 dB and also delayed by 226 ns. We will use this larger delayed signal for calculating the effects.
- If a digital datastream is being transmitted, one effect of this "microreflection" is that if a pulse is transmitted, an echo of that pulse will create a voltage amplitude uncertainty of about 2.6% in the signal 226 ns later.
- The amplitude response will have a "ripple" versus frequency. The frequency of the ripple will be 4.42 MHz (meaning a full cycle plus in every 6-MHz channel), and the amplitude will be 0.45 dB peak-to-peak.
- The group delay will also have a variation versus frequency at the "ripple" rate. The delay variation will be 11.7 ns peak-to-peak.

Many cable operators, on the other hand, do *not* terminate unused tap ports, but only cover them to prevent undue weathering of the connectors. Their reasoning has been that a loose type F terminator acts as both a radiator

and signal pickup point, reducing the shielding integrity of the system. The consequences for reflected signals, however, can be serious. In the case of a terminating 4-dB tap (actually just a two-way splitter), for instance, the return loss when unterminated is typically < 5 dB. If the preceding calculations are made assuming a 5-dB return loss, the delayed signal will be only 20.2 dB below the desired signal, causing the delayed voltage uncertainty to increase to 9.8%, the amplitude response variation to increase to 1.7 dB, and the group delay variation to increase to 45 ns. Although none of the results from a properly terminated system would significantly degrade most signal formats, the second set of values is large enough to be of concern.

If we do a similar calculation at 10 MHz in the upstream band, the principal differences are that the cable loss is much less, and tap directivity, and sometimes return loss, is degraded. Figures 15.5 and 15.6 show the amplitude and group delay variations for some typical situations based on an average of several manufacturers' component specifications.

Though the preceding calculations are simplified to a single reflection, in actual systems, many reflections occur. Since a tapped distribution line can easily contain a bridger, 2 line extenders, and 30 taps, many microreflections exist simultaneously and will add vectorially. Consider, for example, four taps equally spaced along a cable that has 0.5 dB of loss in each segment (a typical upstream value), as illustrated in Figure 15.7. Assume that each has a return loss of 16 dB at all ports and, for simplicity, that each of the first three has a through loss of 1 dB (including any mismatch losses). We can calculate the worst-case effective return loss and VSWR at the input to the first tap as follows (assume the last tap is a terminating type):

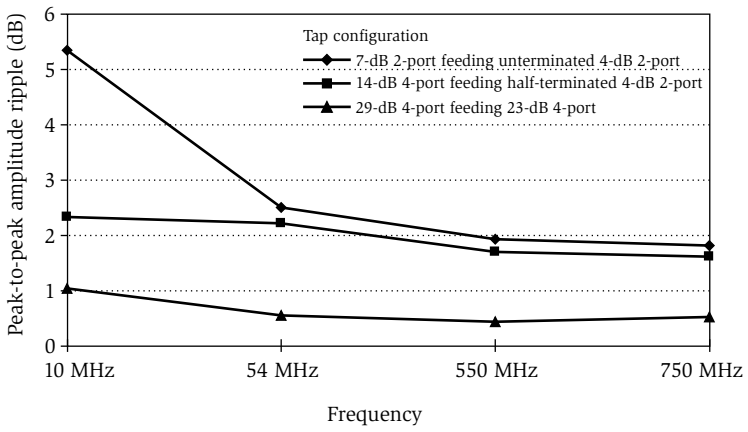


Figure 15.5 Passband frequency response as a function of tapped line configuration.

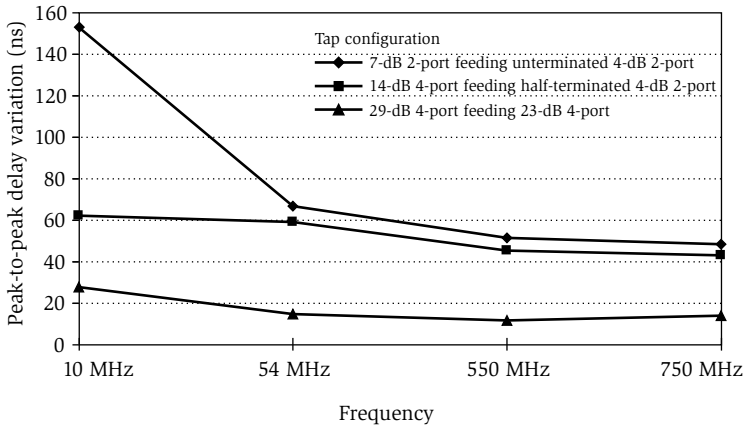


Figure 15.6 Group delay as a function of tapped line configuration.

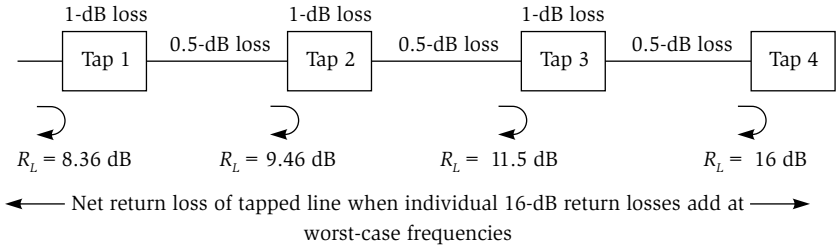


Figure 15.7 Cascaded reflections.

- The signal reflected from tap 4 is 16 dB below the incident signal and at the input port.
- Observed from the input of tap 3, tap 4's return loss will look 3 dB better because the incident and reflected signals are both attenuated by the through loss of tap 3 and the intervening cable loss. Using Equation (15.10), this corresponds to a VSWR of 1.253.
- The 16-dB independent return loss of tap 3 corresponds to a VSWR of 1.376.
- The two mismatches will combine, in the worst case, to the product of the individual VSWRs, or $1.253(1.376) = 1.724$.
- This can be converted back to an equivalent return loss of 11.5 dB using the reverse of Equation (15.10):

$$\text{return loss (dB)} = -20\log\left[\frac{VSWR-1}{VSWR+1}\right] \tag{15.17}$$

This, in turn, is attenuated so that, when observed from the input of tap 2, it is 14.5 dB. Converting to VSWR and combining with tap 2's own input match, we get a worst-case combined VSWR of 2.014 and return loss of 9.46 dB.

Repeating the process at the input to tap 1, we get a net VSWR of 2.24 and an equivalent return loss of 8.36 dB. If the string of components were longer, the equivalent match would continue to degrade though at a slower rate.

Since the phase shift through the interconnecting cables changes quickly as a function of frequency (rotating through a full 360° at the ripple frequency), the reflections from two mismatched components should be assumed to reinforce at many frequencies. This is the mechanism that causes structural return loss in coaxial cable to have sharp peaks at many equally spaced frequencies if the manufacturing process produces equally spaced physical defects. The reflections from multiple mismatches, however, will be less likely to reinforce if the spacings are random: a powerful argument against equally spaced similar components in coaxial distribution lines.

Reflections in Drops

As with the trunk and distribution plant, microreflections in the drop system lead to both group delay and frequency response variations. Equations (15.16) and (15.13) can be used to estimate the magnitude of these effects, whereas Equation (15.12) will provide the ripple frequency when applied to the typical drop configuration shown in Figure 11.5 as follows:

- For R in the equations, use the sum of the splitter isolation plus twice the loss of the cable from the splitter to an outlet.
- For L in the equations, use twice the length of the cable from the splitter to an outlet.
- For V_P in the equations, use the relative velocity of propagation in the drop cable (typically 0.85).

For example, at 550 MHz, a splitter isolation of 15 dB combined with a nonterminated 50-foot long, size 59 drop cable (one-way loss about 3 dB) will create a sinusoidal group delay variation of 21.5 ns peak-to-peak with a period of 8.4 MHz. The corresponding amplitude response variation will be 1.6 dB peak-to-peak with the same periodicity.

Group Delay Due to Diplex Filters

In addition to the group delay caused by reflections, frequencies close to crossover frequency of diplex filters used in amplifiers to separate upstream and

downstream frequency signals will be subject to additional group delay. As discussed in Chapter 10 (see Figure 10.12), typical chroma delay (differential delay between the luminance and chrominance frequencies) per amplifier is 21 ns at channel 2, 9 ns at channel 3, and drops off rapidly at higher channels. Since this delay is cumulative through multiple stages, a typical five-amplifier HFC cascade will have a total chroma delay of 105 ns at channel 2.

Older all-coaxial cable systems and systems using longer cascades typically limit the upstream frequency range to 5–30 MHz so that the spacing between upstream and downstream frequencies is greater and the group delay per amplifier is reduced accordingly.

As a general rule, the total channel 2 group delay through most systems will not exceed 100–150 ns and will meet the FCC's 170-ns requirement. Upstream group delays due to multiple reflections are often more severe than downstream due to the lowered cable loss and, frequently, worse component matches at the lowest frequencies. Upstream transmission is discussed in detail in Chapter 16.

Other In-Band Frequency Response Variations

In addition to response variations at the ripple frequency associated with multiple reflections, channels will experience in-band variations from three causes: the intentional gain/frequency slope of amplifiers (typically 10 dB across the full downstream spectrum), the variation in cable loss with frequency (in both distribution and drop cables), and the precision with which the various loss variations can be compensated within each amplifier (P/V).

In addition to the other effects discussed, individual drops may be subject to additional amplitude and group delay degradation when filters and traps are inserted to tailor the spectrum delivered to customers. Various combinations of high-pass, low-pass, bandstop, and bandpass filters are often used for this purpose, as discussed in Chapter 21.

It is rare for the total in-band response variation within any single 6-MHz channel to exceed 1.0 dB pk–pk in a well-maintained cable system, not including the effects of drop filters.

15.5.8 Local Oscillator and Other Interfering Signals from Receivers

The effects of the performance of consumers' receivers that are directly connected to the network are covered in detail in Chapter 24. In summary, however, local oscillator signals transmitted out the antenna terminals of receivers may have amplitudes as high as −10 dBmV, but are more likely to be −20 dBmV or lower.[5] These signals may be coupled into the downstream signal paths feeding other receivers via the imperfect isolation between tap or splitter output ports. Assuming 20 dB of isolation in such devices and short interconnecting cables, the likely maximum undesired carrier levels at the input to an adjacent receiver

is about −40 dBmV, or 40 dB below the minimum acceptable analog desired signal.

The probability of such interference to any given channel, however, is slight. With normal television design, it occurs only when the receiver generating the interference is tuned exactly seven channels lower than the affected receiver, and is connected directly to the cable drop (as opposed to through a converter). When it does occur, the interfering carrier will appear 5 MHz above the lower band edge or 3.75 MHz above the desired luminance signal and very close to the chrominance signal, where the threshold of visibility is great, as discussed in Chapter 2.

Although the interfering signal magnitude is higher relative to depressed-level, digitally modulated signals, when those signals are placed above the analog signals in the downstream spectrum, the potential for interference exists only in the first seven nonvideo channels above the highest analog channel, at least until digital standards are developed to the degree that digital descramblers are incorporated into consumer-owned receivers.

15.5.9 Antenna-Conducted Ingress from Receivers

Another potential source of interference is over-the-air signals that are picked up by the internal wiring of consumer receivers, by poorly shielded or faulty drop components (including cable and connectors), or coupled across antenna selector switches owing to finite isolation between ports. These mechanisms are discussed in Chapter 24. Such signals can affect the reception at the inadequately shielded receiver, but they can also be conducted out of the antenna terminals and back up the drop cable (thus, "antenna-conducted"). Assuming drops are properly installed using quality components, the largest effect is due to the finite shielding of consumer receivers.

If the source of the off-air signal is a VHF television broadcast station, then the interfering signal will be close in frequency to the corresponding ANSI/EIA-542 Standard scheme channel. Often, in fact, cable channels are phase locked to the corresponding strong off-air VHF stations to reduce the visibility of co-channel interference by eliminating the difference frequency "beat" lines that otherwise occur. A major disadvantage of IRC and HRC channelization plans is that this mechanism for improving subjective picture quality in the presence of ingress interference is unavailable since the visual carriers in each case are phase locked to a master headend frequency source. The potential for interference is especially severe in HRC systems, where the difference between off-air and cable visual carriers is 1.25 MHz and results in a highly visible diagonal line pattern on viewers' television screens when the interfering carrier levels are suppressed by less than about 55 dB below the desired signal.

If the interfering signal is a UHF television broadcast station, then its visual carrier will fall 3.25 MHz above the lower channel boundary of the overlapping Standard cable channel. (Cable channels 73–128 overlap the spectrum utilized for UHF broadcasting.) To the extent that other external radiators cause

interference, the frequencies will differ. For instance, cable channel 19 is often affected by such services as paging transmitters.

Though the probability of destructive interference from external ingress is still relatively low, the interference, when it occurs, is likely to persist since it is not related to the tuning of the inadequately shielded receiver. Tests (discussed in Chapter 24) conducted on typical receivers that were immersed in an external field of 100 mV/m showed that most antenna-conducted egress levels from this cause were −20 dBmV or less.[6] Thus, like local oscillator interference, the probable maximum levels at adjacent receivers exposed to this field strength are about −40 dBmV. It must be pointed out, however, that significant areas in the country are exposed to much higher field strengths.[7] In some cases, well-shielded converters are connected between drop outlets to serve the dual purpose of converting all channels off the off-air frequency in the downstream direction while isolating the cable system from antenna-conducted receiver egress in the reverse direction.

15.5.10 Signal Levels and Stability

Cable systems typically have little problem meeting the various signal level, spectral flatness, and stability requirements imposed by the FCC's rules. In part this is because it would be difficult to meet the noise and distortion requirements if signal levels varied far from optimum levels. It is unlikely, for instance, that total variation across the delivered downstream spectrum will exceed 10 dB while individual carrier levels will probably not vary by more than 5 dB with time and temperature.

The implications of the expected levels and their variation of required video receiver performance are as follows:

- The receiver front end may be exposed to as high as +35 dBmV total RF power across the downstream spectrum (100 channels at +15 dBmV), but more probably 5–10 dB less than that. Additionally, the receiver may be exposed to individual signals as high as +40 dBmV in the return band.
- The receiver must tolerate signals appearing at the image frequency (assuming a standard 41–47 MHz IF band) that exceed the desired signal level by 10 dB.
- The receiver must tolerate adjacent video channels whose levels exceed the desired channel by 3 dB.
- The level of the desired signal could be as low as 0 dBmV or as high as +20 dBmV but will likely not vary by more than 5 dB over time (the FCC allows 8 dB).

The requirements for nonvideo receivers are similar with regard to the total RF power presented at the input port and the stability of the desired signal; how-

ever, the desired signal will be lower by the degree of suppression of carrier levels with respect to analog video signals at the headend. Whether image frequencies are present at the receiver input will depend on the design of the receiver and the frequency of the data carrier.

15.6 Summary

In summary, typical HFC broadband distribution networks provide a broadband thermal noise floor at a level of approximately −45 dBc relative to normal video levels, as measured at the entry point to buildings and referenced to a 4-MHz bandwidth. The noise level in other bandwidths, after set top terminals, or relative to other signal levels is easily calculated.

Various signals may occur adjacent to or within standard 6-MHz channels as a result of network nonlinearities, normal signal loading, and/or attached subscriber receivers. These are summarized in Figure 15.8. For reference, the approximate level of suppression required to maintain a subjectively good analog NTSC picture is shown as a dotted line (see Chapter 2 for a detailed treatment of subjective picture impairments due to interfering carriers).

Finally, downstream signals transmitted through the network will be subjected to the following:

- Hum modulation that may be as high as 3% pk–pk, but is unlikely to exceed 1.5%.
- Microreflections that could be as high as −13 dBc, but are unlikely to exceed −20 dBc, with delays ranging from 0 to 500 ns, with smaller echoes having delays as long as 5 μs.
- Sinusoidally varying group delay (as a function of frequency) with an amplitude as high as 100 ns pk–pk, but unlikely to exceed 30–50 ns. The periodicity of the delay variation will typically be in the 2–5-MHz range. Additionally, channel 2 (in a standard low-split system) may be subjected to chroma delay as high as 170 ns, but more likely 100–150 ns owing to the diplex filters. Finally, individual subscriber drops may experience additional group delay due to the effects of any installed filters.
- In-band frequency response variations arising from microreflections, differential cable losses, intentional amplifier response slope, and imperfect frequency compensation. The total variation from all these causes is unlikely to exceed 1 dB pk–pk in any 6-MHz channel. As with group delay, however, the use of drop filters may add considerable response variations to some channels.
- Received signal levels (for signals launched at video-equivalent levels) ranging from 0 to +20 dBmV at subscriber outlets, with variations within that total range as large as 8 dB (but more likely less than 5 dB) over time and temperature.

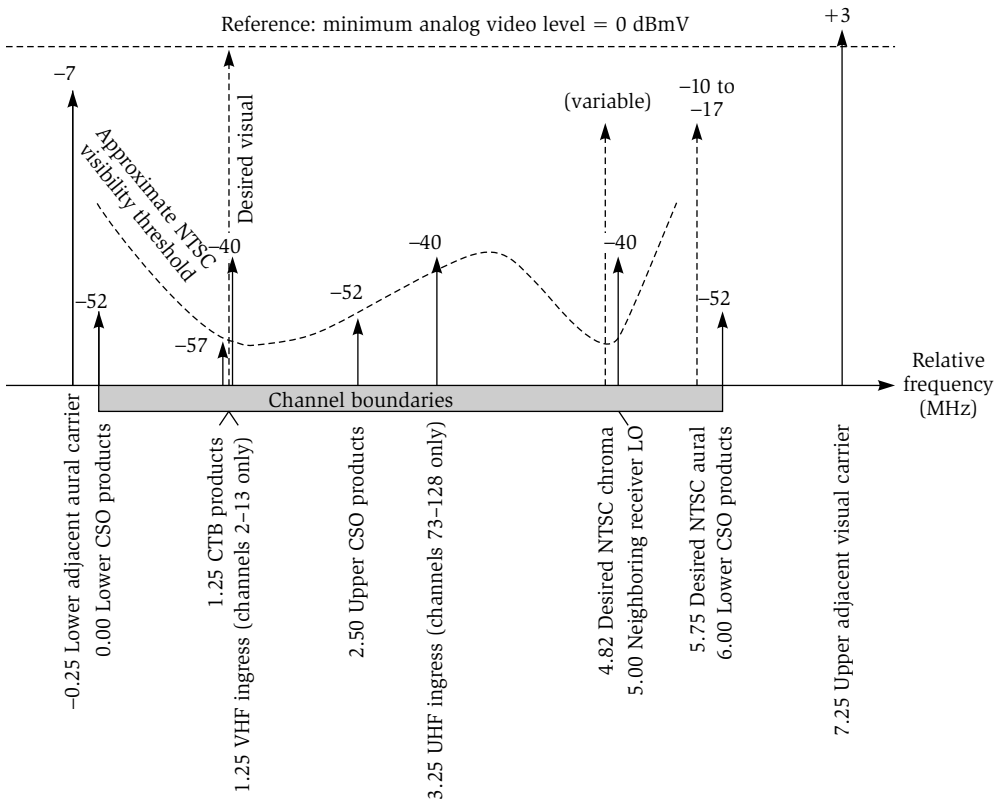


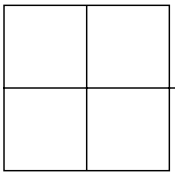
Figure 15.8 Maximum expected levels of extraneous signals in or near ANSI/EIA-542 Standard television channels as received at the input ports of terminal equipment.

As will be seen in the next chapter, signals transmitted in the reverse direction (from subscriber toward the headend) are subject to channel conditions that are less well defined.

Endnotes

1. Hans Stekle (Hrsg.), *Breitbandverteilnetze der Deutschen Bundespost*. Heidelberg: R. v. Decker, 1988.
2. *Code of Federal Regulations*, *Section 47*, available from the U.S. Government Printing Office, Washington, DC, in several volumes covering the entire FCC set of regulations, Parts 0 through 80.

3. *EIA-IS-23 RF Interface Specification for Television Receiving Devices and Cable Television Systems*, unreleased draft document of the Consumer Electronics Association, Arlington, VA, May 12, 1998.
4. Private correspondence from Archer Taylor of the Strategis Group, who doesn't claim to have invented it. We have modified Archer's suggested notation to add a variable cascade factor to broaden its applicability.
5. *Customer Premises Equipment Performance and Compatibility Testing*, CableLabs, Louisville, CO (undated, but published in late 1993). Part 4: Receiver Performance.
6. Ibid.
7. Ibid., Part 2, Potential Impact of Direct Pickup Interference. The report concludes that about 20% of U.S. households will be exposed to external field strengths of at least 100 mV/m from at least one VHF television broadcast station, while 41% will be exposed to that magnitude of signal from at least one UHF station operating at or below 550 MHz. Television receivers, however, are usually placed within buildings that provide some attenuation of the external signal.



Chapter 16
Upstream Issues

16.1 Introduction

This chapter covers a number of topics of interest to those using return plant or needing to understand the issues involved in the operation of return plant. You may wish to review Chapters 10 through 13, which cover relevant topics in distribution systems. This chapter builds on the material in those chapters.

Return, or upstream, signal levels must be managed differently from downstream signal levels, and this topic is covered in some depth. Return laser characteristics are covered, along with methods being used to determine the proper level at which to operate return lasers. Finally, noise on the return path is dealt with. Several methods of analyzing the problem are presented, as are several alternative countermeasures.

Limited two-way cable plant has been built since the mid- to late 1970s. As early as 1972, the FCC required that plant be built to be "two-way ready." It was not until the mid-1990s, however, that the technology began to enjoy adequate applications to justify widespread deployment. Previously, the only applications were provision of a return path for impulse-pay-per-view (IPPV) set-top terminals, some of which used an RF return path, and a few status-monitoring applications. Return analog video was and is practiced, but most links involve trunk runs only and are limited to local backhaul applications. Beginning in the mid-1990s, these applications were joined by HFC-based high-speed data transmission and telephony. More sophisticated status monitoring was deployed in the plant though its required data rates remain modest. Interactive video continues to be a possible future application, though as of this writing, it has not enjoyed widespread acceptance.

The enabling technology that made the widespread use of the return path possible was the introduction of smaller nodal architecture in place of the older tree-and-branch architecture. This made feasible the control of noise buildup and also made it possible to realize adequate bandwidth to support marketable services. Thus, by the mid-1990s, the technology and applications had both arrived to make return plant practical and economically justifiable.

16.2 The Two-Way Node

In early fiber-optic deployments, the device that received optical signals from the headend and converted them to electrical form was called a *receiver*. As two-way services developed, a return optical transmitter was added to the optical receiver. Use of the term *receiver* to describe the device became confusing, so the device came to be known as a *node*. Nodes may be mounted on strand or in a pedestal.

Figure 16.1 illustrates an optical node. It consists of a forward (downstream) optical receiver, one or more reverse (upstream) transmitter(s), and diplex filters to separate signals moving in the two directions on the coaxial cable. The diplex filter, or *diplexer*, is a matched set of low-pass and high-pass filters, which separate signals by direction (due to their difference in frequency). Diplex filters are described in Chapter 10. Also shown is an optional postdiplexer input, which

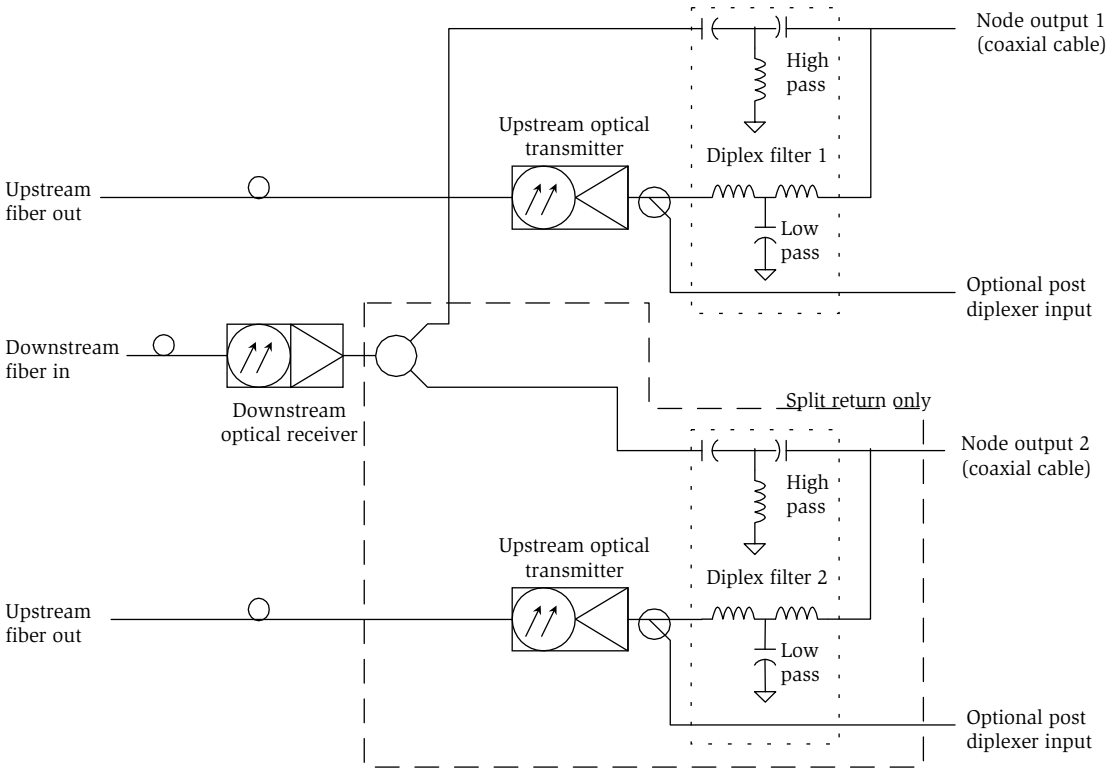


Figure 16.1 Optical node with optional return split.

may be used when a block converter translates multiple input paths to different frequencies. This technique is discussed later.

Nodes typically have one downstream path, which is split internally to several output paths, each serving a portion of the total subscribers served from the node. The upstream paths from all subscribers may be combined to send all upstream signals through the same path. Alternatively, multiple return paths may serve the same downstream path, as shown in the figure. Two node outputs are shown, both served by the same downstream optical receiver but by separate upstream optical transmitters. Splitting the upstream path is one technique used to improve the effective upstream bandwidth of, for example, DOCSIS modems. Since the upstream band is so much narrower than the downstream path, much less upstream capability is provided as compared with the downstream capability.

If more upstream bandwidth is needed, the upstream direction may be split as shown, using two or more upstream optical transmitters. Each transmitter may transmit on a separate fiber strand as shown, or coarse wave division multiplexing (CWDM) may be used to place all upstream paths on the same fiber strand. Normally, CWDM will be much more expensive than using separate fibers. The exception would be if new cable must be installed. That can be more expensive than using CWDM. Other alternatives for improving upstream bandwidth are covered later in this chapter; they include block conversion and digitization of the upstream signals.

High-end nodes, often called *scalable* nodes, not only allow the upstream to be split, but also allow multiple downstream paths to be added if the demand for them develops. This might be needed if a serving area develops a big appetite for pay-per-view (PPV) programming, which requires individual downstream bandwidth.

16.3 Downstream and Upstream Frequency Partitioning

A number of different plans are used in assigning downstream and upstream frequencies. In North America, the downstream band usually begins at 54 MHz, the low end of channel 2 (see Chapter 9). However, some operators prefer to use the frequency band from 50, or 52, to 54 MHz for narrowband services, and so use diplex filters that cut off below 50 MHz. The maximum downstream frequency is usually between 450 and 870 MHz, depending on when the system was built and the economic viability of various bandwidths in a particular system.

The traditional upstream band was 5 to 30 MHz, but recently, the maximum frequency of the upstream band has been increased to either 40, 41, or 42 MHz in order to allow more usable return bandwidth. Caution is advised when using frequencies above 40 MHz, and especially when using frequencies above 41 MHz. The common IF band used by North American television receivers is 41–47 MHz. Not all television sets exhibit good IF rejection to signals entering the antenna terminals because this has not been a requirement. Since return

signals cannot be completely isolated from television receivers, the possibility for interference is significant. Issues relating to subscriber susceptibility are covered in Chapter 23.

A common practice is to specify two numbers to indicate the high end of the upstream band and the low end of the downstream band. A partitioning having a maximum upstream frequency of 40 MHz and a minimum downstream frequency of 54 MHz would be called a 40/54 split. Other splits in common use include 55/70 (Japan) and 65/82 (Australia, New Zealand). Various European systems have used 26/45, 33/50, 42/54, 50/70, and 65/85. Splits tend to be somewhat customized. In certain applications, the split tends to yield more symmetrical spectrum, with larger upstream bands. These are not popular for residential service due to the need to place TV signals beginning at channel 2.

Useful upstream bandwidth may be somewhat less than the entire return bandwidth. On the low end, ingress often is the limiting factor below 15 or 20 MHz. The subject is covered in more detail later. As you approach the upper end of the return band, group delay increases owing to the characteristics of the diplex filters. Whether the group delay is a problem or not depends on the number of diplex filters in cascade, the rate of cutoff of the diplex filter, and the type of modulation used. Thus, the usable bandwidth of the return plant may be lower than the reverse passband.

16.4 Group Delay of Diplex Filters

Of concern with two-way plant is the group delay of the diplex filters used to separate the upstream and downstream frequency bands. The narrower the crossover region, the worse the problem becomes. In the upstream direction, the effect is to introduce group delay to the data signals that are typically carried in that direction. Group delay can produce intersymbol interference, the error in a signaling symbol produced by adjacent data symbols. If bad enough, the bit error rate of the signal is adversely affected. In the upstream direction, it is generally not feasible to equalize for group delay, as is done when necessary in the downstream direction, because the equalization would have to be different for each of several transmitters that may be using the same upstream bandwidth. Group delay is discussed in more detail in Chapter 10.

16.4.1 Upstream Group Delay

Figure 16.2 illustrates the upstream group delay of an amplifier having a 42/52 split.[1] The left vertical scale shows the absolute group delay (defined in Equation (10.35)), and the right scale is the differential group delay in a 1-MHz frequency span. (If we were concerned about carrying NTSC video, we would measure the differential group delay over a frequency span of 3.58 MHz, and we would call it chroma/luma delay.) Notice that at the high end of the return path, the

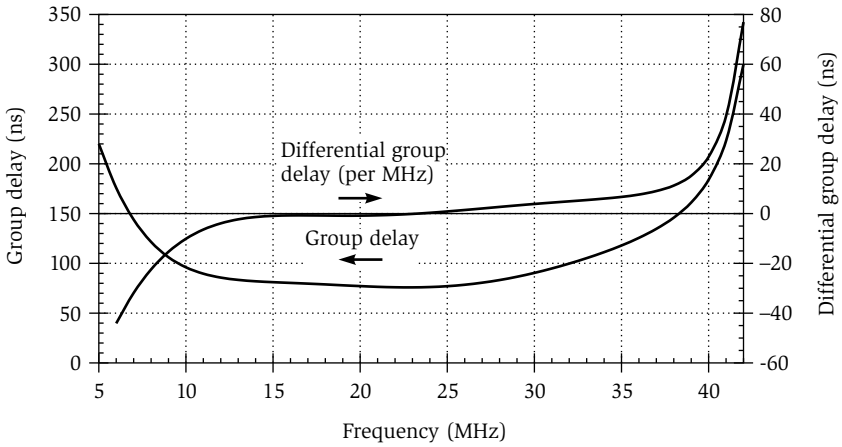


Figure 16.2 Upstream group delay due to diplex filter.

differential group delay increases significantly, as the frequency is approaching the low-pass cutoff of the diplexer. If the differential group delay is too great, then some spectrum at the high end of the band may not be usable. "Too much" group delay is a function of the modulation type and the particular system being used. Higher-density modulation systems (that is, those having a higher bandwidth efficiency) are more susceptible to group delay and other impairments than are lower-density modulation systems.

The group delay increase at the low end of the band is a result of a high-pass filter included in the amplifier measured. The high-pass filter is advantageous in that noise below the 5-MHz low cutoff is not as much of a factor in loading the return system.

Notice that the chart in Figure 16.2 reflects the group delay per amplifier. Multiply the group delay shown by the number of amplifiers in cascade. Optical nodes will exhibit somewhat less group delay because they have only one diplex filter (see Figure 16.1). Figures 10.15 through 10.17 show that two diplex filters are cascaded in an amplifier. Furthermore, amplifiers need more crossover attenuation than do nodes to avoid the stability issue described in Chapter 10.

16.5 Block Conversion

It is sometimes desirable to increase the return capacity of a node. The limited return bandwidth that can practically be used limits the number of services that can be accommodated in the reverse direction. Also, noise buildup may limit the

performance of the upstream bandwidth. Capacity may be increased, and noise buildup reduced, by splitting the return into several sectors (frequently based on multiple cables branching from the node) and using a separate return transmitter for each. Rather than using multiple return transmitters, block conversion has been used as an alternative. A block conversion system takes several return paths and converts all except one to a unique block of frequencies. The blocks are combined and transmitted to the headend using one return optical transmitter.

Figure 16.3 illustrates block conversion in a common configuration that uses a 200-MHz return optical transmitter. Typically, this circuitry would be housed in a node though it might be housed in an ancillary housing. Up to four return paths may be accommodated. The first is coupled directly to the optical transmitter. The other three are up-converted to other frequencies up to about 200 MHz and combined with the unconverted spectrum to supply signals to a return transmitter. At the headend, the process is reversed, developing four individual 5- to 40-MHz spectra to be supplied individually to receivers. Alternatively, it is possible to build return receivers that tune to 200 MHz, eliminating the need for downconversion in the headend.

Single conversion is used for economy. For other applications, double conversion might be used if it is necessary to put many blocks close together, but the

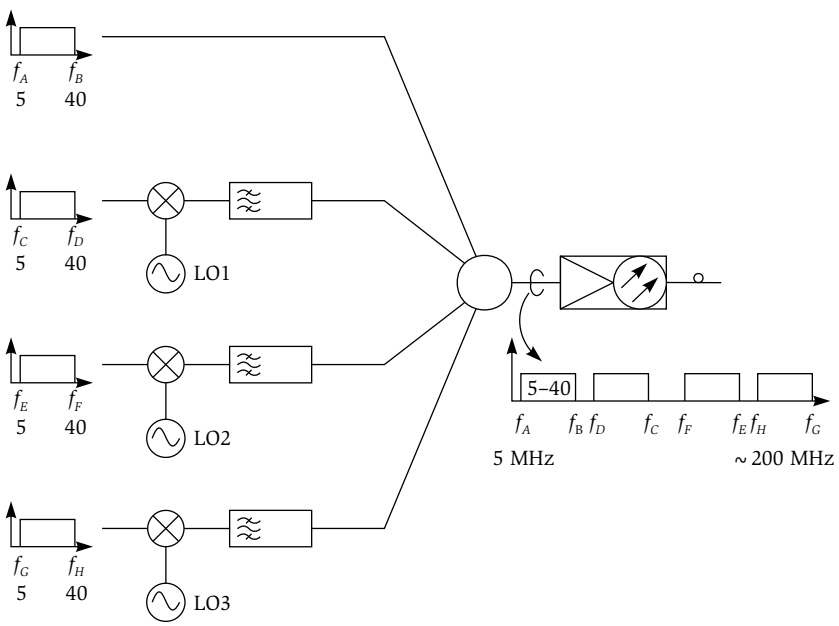


Figure 16.3 Block conversion at a node.

cost can be higher. The single conversion results in inversion of the spectra of the three converted bands so that what was the low end of the band becomes the high end of the converted band. The conversion back in the headend will reinvert the spectrum. Inversion is brought about by the practical need to use high-side local oscillators.

In some cases, it is necessary to provide very accurate frequency conversion in the node and headend, such that the frequency to which a signal is restored in the headend is very close to the frequency at which it existed before conversion at the node. If this is required, then most commonly, the oscillators in the node will be phase locked to those in the headend, or vice versa. A pilot tone is transmitted between the node and headend to allow phase locking.

Block conversion increases the bandwidth available per customer and reduces noise funneling without requiring more optical transmitters and receivers. However, since the one optical transmitter is loaded with more signals, the performance of that transmitter must increase, or the OMI must be reduced. This reduces the carrier-to-noise ratio available in the optical portion of the return path. The use of block conversion may force an operator to use a DFB return laser rather than a lower-cost F–P laser.

16.6 Return Signal Level Issues

Of paramount importance to the operation of the return path is the proper handling of signal levels. The issue is far more complex in the upstream direction than it is in the downstream direction. In the upstream direction, you deal with many transmitters located on or in subscribers' homes, with each signal path having a different gain between the home and the headend. Several conditions must be addressed in setting the level of the transmitters. It is as important to design and balance the reverse path as it is the forward. If you balance the forward plant and let the reverse fall where it may, reliable return performance will not be obtained.

16.6.1 Measurement of Return Signal Levels

In Chapter 4, the definition of "signal level" of a digital signal was presented as the level, expressed in dBmV, at which a sine wave would produce the same heating in a resistor as does the signal being measured. That is, the level of a digital signal is taken to be the average power. Most digital signals have the characteristic that the amplitude of the signal changes for different transmitted states, so the peak level is greater than this average. The peak-to-average ratio is small with QPSK, and indeed, would be zero were it not for transmitter filtering. Higher-order modulation exhibits a greater peak-to-average ratio, as shown in Chapter 4. Peak-to-average ratio is rather difficult to measure, so the *NCTA Recommended Practices for Measurements on Cable Television Systems*,

"Upstream Transport Issues" section, recommends that each manufacturer state the peak-to-average ratio of his or her modulation.[2]

Many signals in the upstream direction employ time division multiple access (TDMA), a technique in which many transmitters transmit on the same frequency at different times. An explanation is presented in Chapter 4. Measurement of individual elements can be accomplished by triggering the measuring instrument (usually a spectrum analyzer) on some reference time, allowing one of the signals in the return "parade" to be identified. This triggering signal would normally be supplied by the system being analyzed. First-generation equipment does not necessarily have the capability to provide such a trigger, however.

If you set a spectrum analyzer to the zero span mode, it can be used to measure the signal level on only one frequency, with the bandwidth of the analyzer determining the bandwidth of the measurement. You may be restricted as to the measurement bandwidth due to adjacent carriers, and this will vastly complicate the measurement.

Figure 16.4 illustrates the measurement issues for a TDMA signal in the return path. Figure 16.4(a) shows the preferred test configuration: received signals are supplied to the using system and also to a spectrum analyzer used to measure signal level. Ideally, the using system would supply a triggering signal to the spectrum analyzer, which would allow identification of the return signal

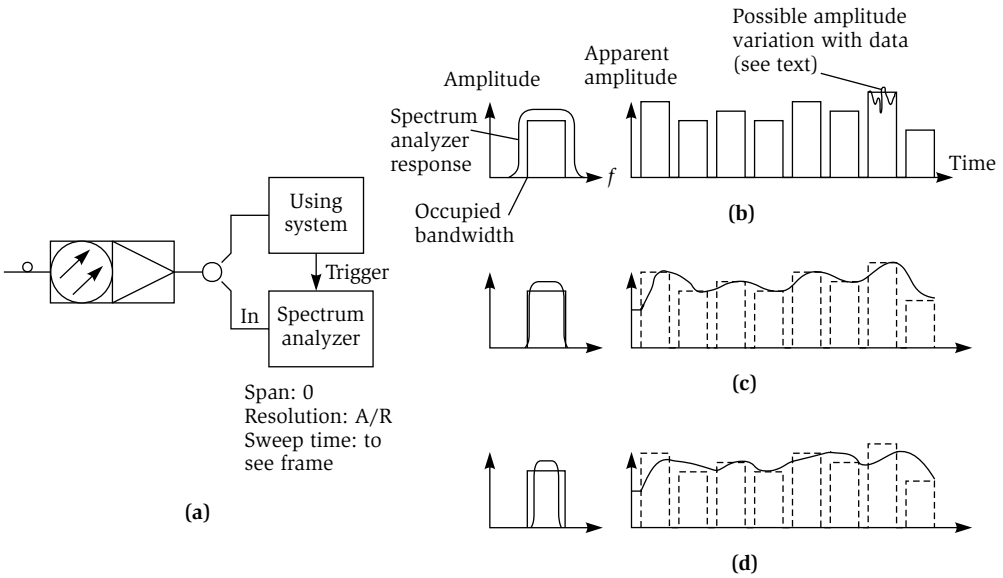


Figure 16.4 Measurement of TDMA return signals. (a) Instrumentation to observe return levels. (b) Wideband measurement of amplitude. (c) Measurement with restricted bandwidth. (d) Measurement with severely restricted bandwidth.

to be observed. Again ideally, the spectrum analyzer would have a trigger input and would have a time base setting that allows separation of the return signals into the individual elements. These ideal situations do not exist with all first-generation hardware, but it is hoped that future systems will allow for this type of observation.

The spectrum analyzer is set to the zero span mode, in which it operates as a fixed tuned receiver, plotting amplitude versus time of the signal, as shown in Figure 16.4(b), (c), and (d) for three different resolution bandwidths of the analyzer. If the resolution (IF) bandwidth of the analyzer is wider than the occupied bandwidth of the signal being measured, as shown at the left of Figure 16.4(b), the waveform might look as illustrated in (b). The TDMA parade consists of a number of responses from different transmitters, each being received at a slightly different level at the headend. Ideally, the long loop automatic level control (ALC) would bring all elements to the same level. However, there will be differences due to the limited resolution with which the transmitters can adjust level.

Issues Regarding Making Accurate Measurements of TDMA Signals

Particularly with higher-density modulation formats (bandwidth efficiency of greater than two bits per hertz), the amplitude will change within one time slot since a number of bits are transmitted in each time slot, and different bit combinations correspond to differing amplitudes. This is illustrated to the extreme right of Figure 16.4(b) and further complicates the job of accurately measuring signal level.

Yet another difficulty is that, if the spectrum analyzer resolution can be set wider than the occupied bandwidth, it will tend to track envelope peaks, measuring the peak, and not the average, signal level. This will necessitate adding another correction term.

Figure 16.4(c) illustrates a situation in which the spectrum analyzer bandwidth is slightly narrower than the occupied bandwidth of the signal. Analogous to measurement of the amplitude of an analog television signal, if the bandwidth is too narrow, the individual packets will not be present long enough to allow the analyzer output to stabilize at the correct level for each packet. Figure 16.4(d) illustrates a further reduction in bandwidth of the analyzer, showing even more difficulty in accurately reading the bandwidth of any element.

One reason that the spectrum analyzer bandwidth may be too narrow for the signal being measured is that the signal may be wider than the widest bandwidth of the analyzer. It is not uncommon to deal with return signals having a 2-MHz occupied bandwidth, and some analyzers are not equipped with IF filters wider than 1 MHz. Also, if adjacent channels are occupied, it may be necessary to reduce the bandwidth of the analyzer in order to avoid interference from those adjacent signal(s).

Some spectrum analyzers use IF filters with relatively high shape factors (the ratio of bandwidth at high attenuation to bandwidth at lower attenuation). Such filters tend to ring less and can allow accurate measurements to be made faster. However, it does mean that the resolution bandwidth may need to be less than the occupied bandwidth in order to reject adjacent channel power. Other analyzers use filter shapes that are more squared (somewhat as illustrated in Figure 16.4). These filters, if available in the correct bandwidth, do a better job of rejecting adjacent channel signals, but they tend to ring more, introducing yet another source of error.

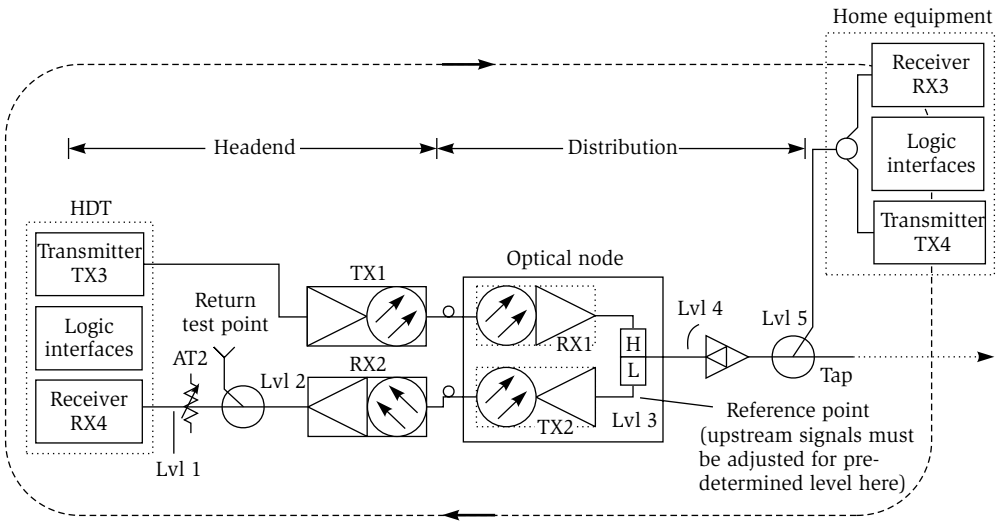
Making the measurement with a significantly narrower spectrum analyzer bandwidth can yield a fairly accurate average channel power measurement *if* return signals are present substantially all the time. This can be useful in checking that a long loop ALC system is set up to operate at the desired point. It cannot tell anything much about the amplitude of a specific element in the return signal. If there are gaps in the received signal, the results can be quite misleading. Consider a telephony system that multiplexes 24 telephone lines into a single TDMA stream. If the node has fewer than 24 lines, there will be some gaps in the received signal. These gaps can provide misleading results when you attempt to measure the signal level of the stream by looking at average power.

In summary, it is difficult to make accurate measurements of the amplitude of TDMA signals in a real network. Gross errors can often be spotted fairly easily, but very accurate measurements are hard to make. The *NCTA Recommended Practices* document, referenced earlier, defines the "level" as being that which would be measured if the signal was on continuously. It further requires the manufacturer of a signal-measuring device to publish a valid procedure to allow a user to translate the level indicated to the defined level. However, this is not always easy to do.

Measurement of the level of return signals using more advanced techniques, such as code division multiplexing and multiple carriers, is not covered here. The vendor of such systems should provide guidance for making the required measurements.

16.6.2 Long Loop Automatic Level Control

Figure 16.5 illustrates the critical points in the return signal level control issue. It is similar to Figure 9.8, which addresses the same issue from the headend. As described in Chapter 9, the return service includes a long loop automatic level control (ALC), wherein the signal level is measured in the headend (Lvl 1 in Figure 16.5), and a signal is sent to the home transmitter, TX4, to adjust its output such that the correct level is received at the headend. This corresponds to some level at the tap, Lvl 5, and to other levels at each RF amplifier (Lvl 4) and at the node, Lvl 3. The point is that, once the system controls the signal level at Lvl 1, all the other signal levels are set according to the design of the system.



An ALC loop is set up between the headend and each home system individually. The headend senses level of the return signal and adjusts the upstream transmitter such that the correct level is received at the headend receiver. The most critical level is that at the reverse optical transmitter, Lvl 3.

Figure 16.5 Pertinent points in return signal level control.

The most critical level is that at the return optical transmitter, Lvl 3. The dynamic range (from maximum signal level down to noise level) tends to be relatively low for return path optical transmitters, both due to technology and cost. Dynamic range is explored in more detail later. See Chapter 12 for a detailed discussion of optical technology.

The system must be set up to optimize the level at Lvl 3. However, Lvl 3 cannot be directly observed by the adjusting system. Instead, the gain between the point marked Lvl 3 and that marked Lvl 1, where the level observation is made, must be known and must be stable. The target value of Lvl 1, to which the system is adjusted, is set based on the desired value of Lvl 3. At the same time, it is necessary to set the gain in the system such that the input and output signal levels at all amplifiers, represented by Lvl 4, are correct. Attenuators and equalizers are often provided in the return amplifiers for this purpose and must be properly selected during system alignment. "Correct" is based on the manufacturer's specified operating levels for the conditions under which the amplifier is being used. When these levels are set (by design and alignment of the system), then the level at the tap, Lvl 5, will be set by the tap value and the gain from there to the node.

A few systems have attempted to use automatic gain control (AGC) in the return path. This is not recommended for a number of reasons. First, it fights the

setting of levels as described earlier. Second, at times it may be normal for the number of carriers in the reverse direction to change. For example, some services employing time division multiple access (TDMA) may permit signal dropout as a normal condition at times. Further, as return services are added and deleted, it is not desirable that gain in the return path change. Next, real return paths often branch and have a different number of amplifiers in each branch. AGC can significantly complicate operation of such systems.

In some cases where reverse AGC has been employed, it has been done with a pilot tone generated somewhere near the end of the system and detected by all amplifiers and the node. This may work, but raises many questions when downstream signals branch after leaving the node. With branches having a different number of amplifiers in cascade, where do you put the pilot tone generator? You would logically place it at the end of one of the branches, but that leaves the other branches with no coverage.

16.6.3 Thermal Gain Control

It is often desirable to provide thermal control of return path equipment. Even at return frequencies, cable exhibits some variation in loss with temperature, and it may be desirable to also compensate for variations in the operating point of lasers. This is often done by using thermistors (temperature-sensitive resistors) to sense temperature level inside the return optical transmitter and adjust gain based on temperature variation. It is also practiced at times to cool the laser, as mentioned in Section 16.7.

16.6.4 Return Signal Levels at the Tap

Of interest to those designing or using systems needing the services of the return path is the issue of the signal level at the tap, Lvl 5, needed in order to produce the desired level at Lvl 3. Intuitively, we might assume that the difference between the lowest and highest levels would be about the same as the difference between the lowest and highest tap value in the system. Tap values are chosen such that the downstream level at the tap is more or less constant, fixed by system design. This calculation is done at the highest frequency of interest, where cable loss is maximum. At the return band, cable loss is much lower, so the level required at the tap, Lvl 5, would be lower at lower tap values. Generally, those lower value taps are located farther from an amplifier (in the downstream direction).

Analysis of the return gain characteristics of real nodes, however, shows that this line of reasoning is oversimplified. Sometimes low-value taps are located at the end of a long cable run, so there is little upstream loss at the low frequencies composing the return band. Other times, though, low-value taps are located after a directional coupler and near an amplifier. At those locations, the loss is primarily flat loss (not a function of frequency), and the required output at Lvl 5 will be higher.

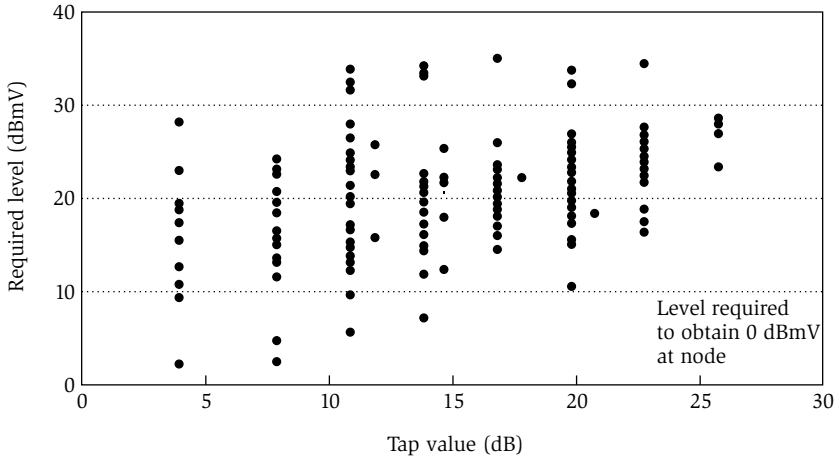


Figure 16.6 Required signal level as a function of tap value.

Figure 16.6, from the analysis of one real node, shows the levels found to be required at the tap (Lvl 5 of Figure 16.5) in order to produce a level of 0 dBmV back at the node (Lvl 3). Each dot represents a tap of the value shown on the *x* axis, which requires the *y*-axis level at any tap port. Notice that low tap values exhibit a wide variation in level required, with the level range tending to compress slightly as you move to higher tap values. This slight compression is a result of having few places to use high-value taps, other than shortly after an amplifier. This data is from one node, but examination of other nodes owned by the same and other cable operators shows the same tendency.

It is not necessarily desirable to have a node that requires low signal levels at the return input. It is desirable to require the highest level that can be supported by the reverse amplifiers that supply signals to the node. This will help minimize the effect of noise entering the plant (primarily from homes). Of course, you must ensure that the requisite signal levels are available at the home transmitter. The use of 0 dBmV as the input to the node is not intended to be a real case but to provide a convenient normalized reference point.

Figure 16.7 illustrates the signal level issue another way. The population of taps having certain level requirements is plotted for the node of Figure 16.6. The distribution of levels appears to be a very crude approximation to a normal distribution. It illustrates that, though there is a clumping of levels, a significant number of taps will require a higher or lower level. The product designer must plan for the required level range, also allowing for variations from system to system, errors, and variations with temperature. Note that the level of the signals at the home will vary because of temperature-dependent changes in gain in the system.

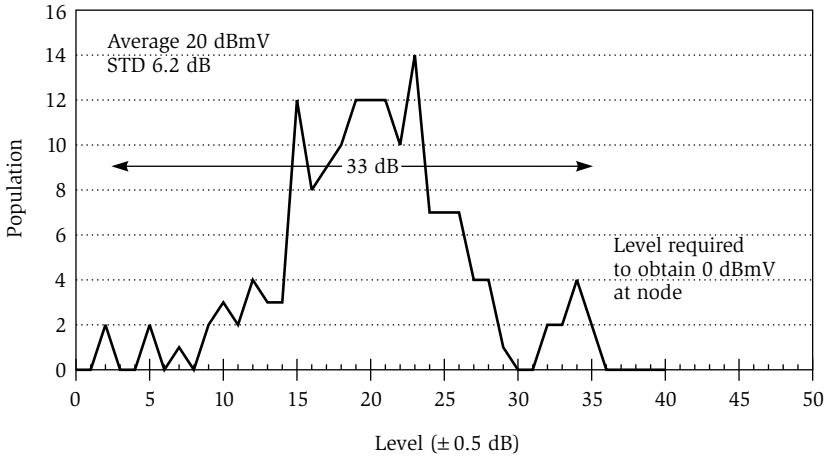


Figure 16.7 Distribution of tap signal levels.

When the signal level received at the headend changes, this will cause the signal level commanded from the home to change, due to operation of the long loop ALC.

Reducing the required tap return level range would be desirable, in that a reduced level range specification would provide some reduction in cost and would enhance performance in at least some cases. Homes with the lowest signal level requirement are also the homes that have the greatest potential to introduce noise into the return path. A wide level range implies high output level capability from the device, which implies possible high noise emission when the level is set low. It also implies rather larger current drain since the output amplifier must have the capability to handle large signals with minimum distortion.

16.7 Optional Ways to Specify Return Lasers

At least four types of return transmitters are in common use at this time: digital, cooled distributed feedback (DFB) lasers, uncooled DFB lasers, and Fabry–Perot (F–P) lasers. The list is in the order of decreasing performance (and decreasing cost). It is required to select the correct laser for each application, bearing in mind the eventual loading, not just the initial loading. Refer also to Chapter 12, which discusses laser technology.

16.7.1 Data- or Video-Grade Classification

An early method of classifying return transmitters (and receivers) is to classify them as to data grade or video grade, with the implication that video-grade lasers

are better. This method has proved unsatisfactory because it doesn't take into account the real requirements placed on the system. A return path with many data signals requires at least as good a laser as does a return path with one analog video signal. Further, there is no standard criteria for classifying lasers: in at least one case, a manufacturer interpreted "data grade" to mean capable of handling one FSK return signal only. When it was used with multiple QPSK return signals, it was incapable of handling the job.

16.7.2 Discrete Carrier Testing and Classification

An older method of measuring the performance of return components is taken directly from downstream measuring practice. Either two or four CW carriers are used to excite the laser, and the results are measured on a spectrum analyzer. The second- and third-order beat products are recorded.

Figure 16.8 illustrates the method. Either two or four oscillators are combined, with identical output levels, to produce a spectrum, usually on the highest T-channel carrier frequencies used by the system. The performance of the system is reported as the composite second- (CSO) and third- (CTB) order beats and the cross modulation, as if this were a miniature downstream system. Typically, lasers tend to be more second-order than third-order limited, so CSO becomes the dominant specification.

The attenuators and bandpass filter are shown for practical reasons. The attenuators help isolate the signal sources so that you don't suffer intermodulation distortion in the output circuits of the generators. The bandpass filter (BPF) tuned to the frequency of measurement prevents overload of the spectrum analyzer.

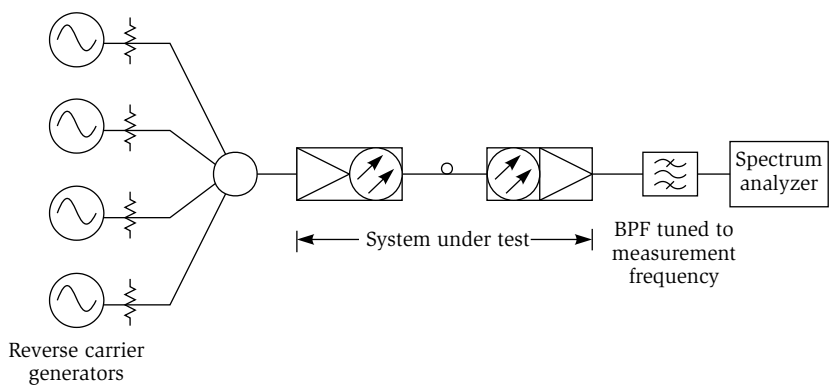


Figure 16.8 Four-tone measurement of a return optical system.

In evaluating the performance of the system for carrying a limited number of signals, some practitioners have tried to predict the frequencies at which the maximum distortion is present. These frequencies are then avoided in order to maximize the performance of the link. The cost of doing this is that many potentially useful frequencies are excluded. You are left with lower capacity and less ability to move carriers to avoid noise. Another problem with this approach is that, as we will show, when the laser goes into clipping, the normal model used for predicting the frequencies of distortion components is invalid. Also, this technique fails to stress the laser the same as it would be stressed with a number of real digital signals. Again, this is explained later.

Yet another problem encountered in comparing specifications between manufacturers by using discrete test tones is that some manufacturers, following downstream practice, quote specifications with equal level carriers. Others assume that some number of the carriers represent analog signals, and others represent digital signals that are carried at a lower level. They reduce the level of some of the signals when making measurements.

Still another problem in comparing specifications is that manufacturers may choose different channels on which to place the carriers. Depending on the placement, differing numbers of composite beat products will exist on the measurement frequencies. Only the most careful reading of the data sheet can reveal the exact conditions under which the measurements were made.

16.7.3 Noise Power Ratio Testing and Classification

A newer method of characterizing systems has become common practice. It comes much closer to testing the laser as it will be operated in service. The technique is not really new: it has been used in the telephone industry for years to characterize linear systems, such as FDM (frequency division multiplex) coax and radio systems. HFC systems are FDM systems because they employ multiple frequencies to carry different information.

Figure 16.9 illustrates the so-called noise power ratio (NPR) testing technique. A broadband white noise generator is used to simulate a spectrum filled with data signals. Data signals generally have a flat spectrum over the occupied bandwidth of the signal (see Chapter 4), so simulating them with a white noise signal is reasonable. (“White noise” simply means that the spectrum is flat over the frequency range of interest.) A bandpass filter limits the spectrum to that for which the system under test is rated.

An attenuator is used to isolate the bandpass filter from a notch filter that follows. The notch filter removes the white noise in some measurement bandwidth. The total power in the signal is important, and would be measured using, ideally, a thermocouple type power meter. The variable attenuator is used to adjust the signal level supplied to the system under test.

A spectrum analyzer, protected if necessary by a bandpass filter tuned to the notch frequency, is used to observe the output. The notched frequency will be

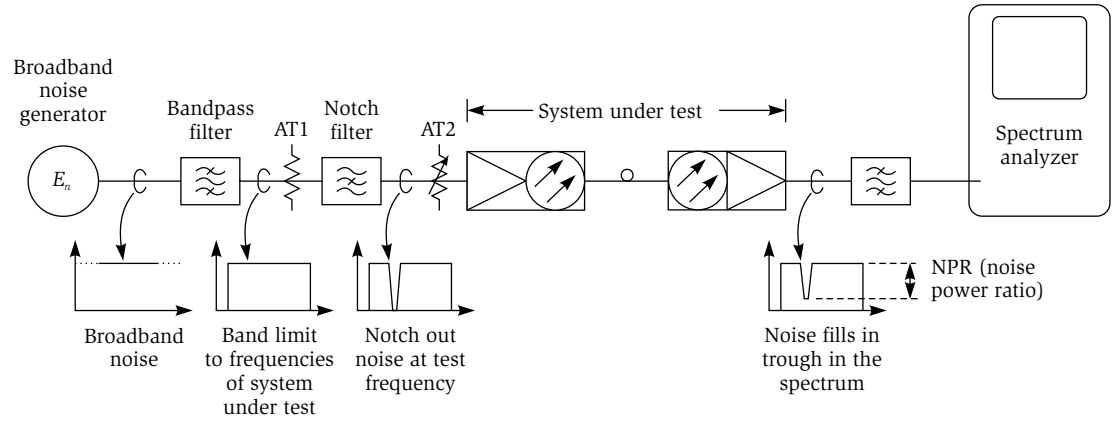


Figure 16.9 Noise power ratio testing of a return path.

somewhat filled in by the system under test. At low signal levels, the filled in noise is due to noise introduced by the system under test. At higher signal levels, it is filled in by distortion introduced by the system under test. It is not possible to look at the filled in level and tell whether the fill is due to noise or distortion or both. The ratio of noise on the flat portion of the spectrum, and that in the notch, is called the noise power ratio (NPR). Some practitioners have called it carrier to noise plus intermodulation noise, or C/(N + IMD).

Use of NPR to Set Operating Level

If we were to plot the NPR against the power supplied to the laser, we would obtain a plot similar to that of Figure 16.10. The total power supplied to the laser is plotted on the *x* axis. Often the quantity plotted will be the power spectral density, which is the total power divided by the bandwidth over which the power exists.

At low signal levels, the NPR is noise limited. One would expect the NPR to increase 1 dB for every decibel increase in total power into the laser. In practice, there will often be a region on the left of the diagram in which the NPR increases faster than decibel for decibel.[3] This is especially true for unisolated distributed feedback (DFB) lasers feeding long fiber-optic cables. The reason is that at low signal levels, lasers are susceptible to reflections from the optical cable, due to Rayleigh backscatter. The reflected light "pulls" the laser wavelength, creating noise. As the drive to the laser increases, the laser tends to "chirp" (change wavelength with instantaneous drive) more, reducing the pulling by reflected light. The reflected light is a function of the length of optical cable used: the

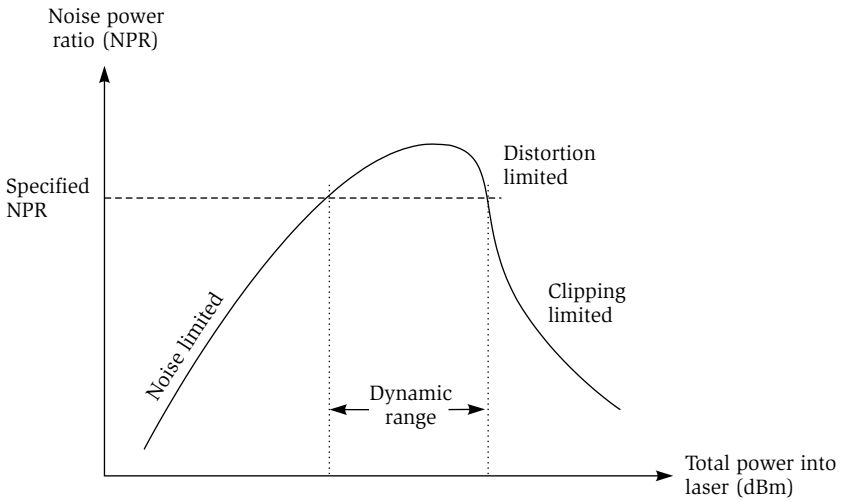


Figure 16.10 NPR as a function of the total power into the laser.

longer the cable, the more reflection. DFB lasers tend to be more susceptible than are F–P lasers because their wavelength is chirped less by modulation, making them more susceptible to wavelength pulling by reflected light. Many modern lasers are isolated and tend to change NPR at about a 1:1 rate with signal level.

Wavelength pulling by reflected light is a phenomenon similar to the tendency of an oscillator to oscillate off of its normal frequency if a signal at a close frequency is supplied to the circuit. The more the separation between the two frequencies, the less the tendency to pull. Chirp refers to the changing of frequency (wavelength) of an oscillator (or light source) with applied modulation.

At some drive level, the NPR is maximized. Above this level, the NPR drops quickly. The initial drop may be due to distortion, primarily second order, but you quickly enter a region in which the primary effect is from laser clipping. Clipping noise tends to be the dominant mechanism by which the NPR is reduced on the right side of the graph, and is covered in greater detail later in this chapter.

NPR should not be confused with carrier-to-noise ratio, C/N. C/N represents the ratio of a signal to the noise lying under that signal. It is a function of signal level and the total noise provided from thermal sources plus the excess noise generated in real electronics. On the other hand, the noise in the notch of an NPR measurement either is a function of the thermal and excess noise or, at higher levels, is caused by distortion, which transfers noise from other frequencies into the notch. The mechanism may be the familiar second- and third-order distortion, or much of the noise filled in the notch may be clipping noise from the laser

(or electronic amplifiers). Do not look for a simple and universal relationship between NPR and C/N. Later in this chapter, we show how to compute C/N from NPR.

One convenient way to establish the level at which to operate the laser is to determine the range of input signal levels across which some specified minimum NPR is obtained, as indicated in Figure 16.10. The laser should be operated with total power within this "dynamic range." The engineer might be tempted to operate the laser at the total power that represents the highest NPR. This will work, but because the dropoff in performance is so steep on the right side, he or she may want to operate with slightly less power (to the left of the maximum NPR). (In some lasers, the slope on the right side has a region of gentler slope, where it is limited by second-order distortion. In other lasers, the slope tends to be quite steep just to the right of the peak. This indicates that the laser performance is limited more by clipping than by second-order distortion.)

A noise power ratio test has several advantages over a four-tone test. Noise has an amplitude probability distribution that is different from that of four carriers, but is similar to that of a number of digital carriers added together. The central limit theorem from statistics states that, when a number of independent variables having any arbitrary distribution are added together, the resultant distribution is normal. A normal distribution also describes the amplitude distribution of noise.

Further, the measurement is independent of the bandwidth of measurement, so long as the bandwidth is significantly less than the notch width. It is also independent of the characteristics of the spectrum analyzer detector, so repeatability is good. The only other instrument (besides a spectrum analyzer) needed is a thermocouple power meter, which is low in cost and very accurate.

Use of BER to Set Operating Level

An alternative way to look at the optimum laser operating point recognizes that the vast majority of signals likely to be carried in the reverse direction use low-density digital modulation. The return spectrum is loaded much as shown in Figure 16.9, except that a modulated (usually QPSK) signal is added into the notched frequency band. The bit error rate is plotted against the signal level.[4]

Figure 16.11 illustrates the principle. At low signal levels, the bit error rate (BER) is poor owing to noise corruption. As the signal level increases, the BER gets better (lower on the *y* axis is better BER). In the central part of the diagram, the BER is so good it is below the chart. Toward the right side of the graph, the BER again begins to deteriorate, due to clipping. The operating total signal level into the laser is set to be somewhere in the middle of the range where the BER is extremely good.

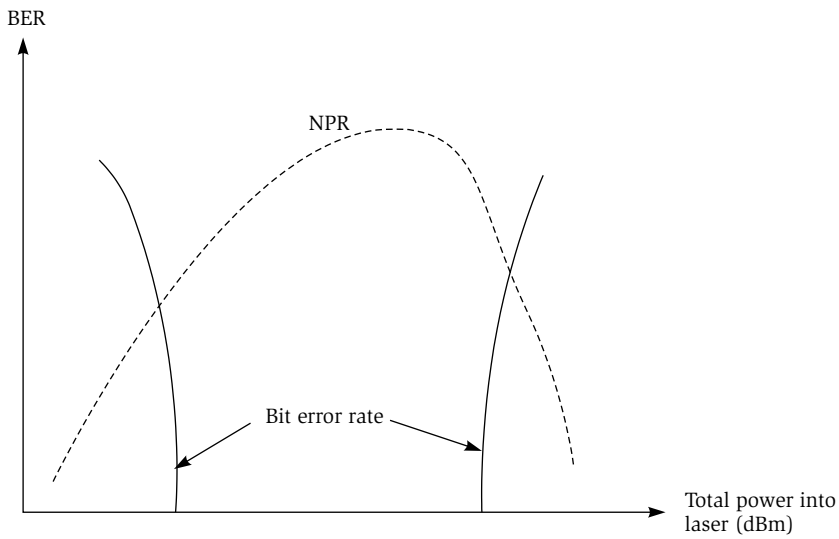


Figure 16.11 BER versus signal level.

16.8 Characteristics of Return Lasers

Two basic types of laser diodes are used in return path service today: Fabry–Perot (F–P), and distributed feedback (DFB) lasers. Chapter 12 provides additional information on lasers used in cable TV work. DFB lasers may be either cooled, with an integral thermoelectric cooler, or uncooled. F–P lasers tend to have lower performance but are lower in cost. Since lasers perform somewhat more poorly at high temperatures, a cooled laser can offer better performance at higher optical output levels. The cooler can also reduce the shift in transfer function of the laser with temperature. However, the cooler increases power consumption and usually increases cost. Electronic circuits have also been developed that compensate for the changes in laser characteristics with temperature. Besides use of analog returns and block conversion, digital return systems have been developed whereby the entire return path is digitized and returned to the headend. At the headend, the signal is turned back into RF form so that signals may be supplied to legacy systems.

16.8.1 Fabry-Perot Lasers

Fabry–Perot lasers have their frequency controlled by the spacing of mirrors at each end of the laser. The frequency control mechanism is such that the laser can oscillate simultaneously, or jump in rapid succession, to several wavelengths that are close to each other. Each wavelength propagates through the fiber at

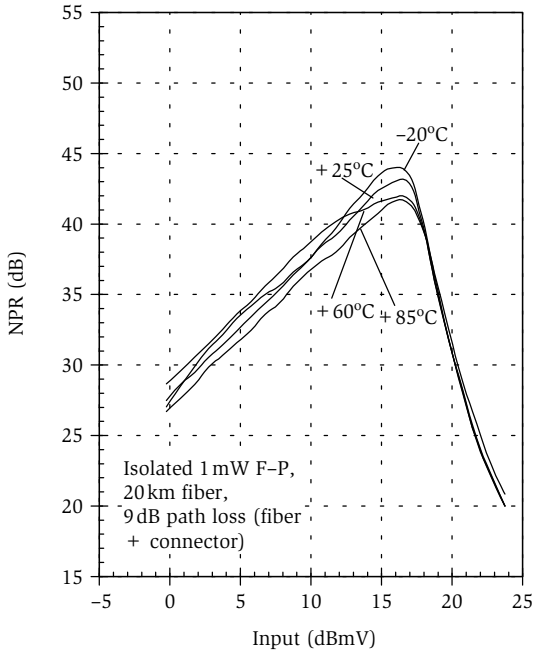


Figure 16.12 NPR performance of an F–P return laser diode.

slightly different velocity due to chromatic dispersion in the fiber.[5] As shown in Chapter 12, this results in restricted performance in terms of carrier-to-noise ratio or NPR on the low side.

Figure 16.12 illustrates the NPR performance of a real F–P laser diode using the NPR test described earlier. On the left side of the graph, the performance is limited by inherent noise in the device. It increases almost precisely 1 dB for every decibel increase in signal level. Above the level of peak NPR performance, the drop in performance is precipitous as the laser rapidly enters the region in which the signal is clipped.

Note that Figure 16.12 is the same plot as Figure 16.10, except that Figure 16.12 is for a real laser. A strategy for setting the level into the laser is to operate near or just to the left of the point of greatest NPR. You would apportion the total power desired, perhaps +15 dBmV for the transmitter shown, over the actual occupied bandwidth of all signals. This subject is covered more fully later in the chapter.

16.8.2 Distributed Feedback Lasers

DFB lasers operate in a similar manner to F–P lasers, except that a diffraction grating restricts the operating wavelength of the laser primarily to a single mode

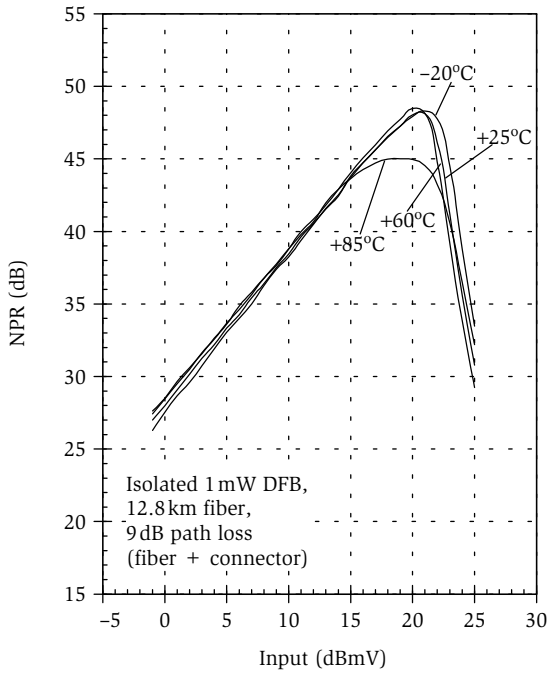


Figure 16.13 NPR performance of a DFB return laser diode.

(wavelength) of oscillation. This means that, relative to an F–P laser, the DFB laser exhibits a lot less noise and distortion as a result of multiple propagation speeds in the fiber.

Figure 16.13 illustrates the NPR of an uncooled DFB return laser diode on the same NPR scale as the F–P laser of Figure 16.12. (The absolute values on the *x* axis are irrelevant and may vary from one transmitter to another.) Both diodes (Figures 16.12 and 16.13) were measured using the same 9-dB path loss. The increase in NPR with increasing signal level on the left side is only slightly greater than decibel for decibel, because the laser is isolated. That is, optical power reflected to the optical transmitter due to Rayleigh backscatter is diverted in an isolator and not allowed to enter the laser. Chapter 12 explains Rayleigh backscatter and the deleterious effect it can have on transmitter operation.

As shown in Figure 16.13, the NPR performance of the laser is somewhat better in terms of carrier-to-noise ratio and dynamic range at room temperature and below than it is at high temperature. For this reason, cooled DFB lasers are sometimes used in return path operation, and they may be necessary where high power levels are required. However, the added cost and power drain (which adds heat to the node) often argue against adding a cooler. The performance of

uncooled DFB lasers is considered adequate for many purposes, though long, heavily loaded paths may benefit from a cooled laser. The laser of Figure 16.13 was embedded in a temperature-compensated transmitter module. Had it not been, the performance range with temperature would have been worse.

Effect of Distance on DFB Link Performance

Figure 16.14 compares the NPR curves for one isolated DFB transmitter with different lengths of fiber, expressed as the loss of the fiber and connectors. Shown are 3-dB through 15-dB links. As expected, with longer fiber lengths the peak NPR curve drops, as a result of lower received signal level. Had the test been done with nonisolated lasers, the performance on the left would have dropped off much faster, as a result of Rayleigh backscatter noise reflecting into the laser.[6]

Digital Return

Yet a third technology for returning signals is available. This is to digitize the return band at the node and to transport the digitized signals to the headend. At

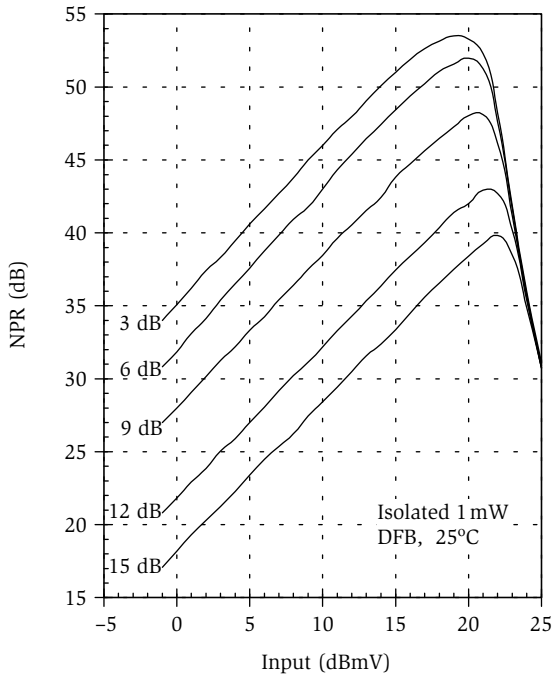


Figure 16.14 NPR Performance of a DFB laser for different return path lengths.

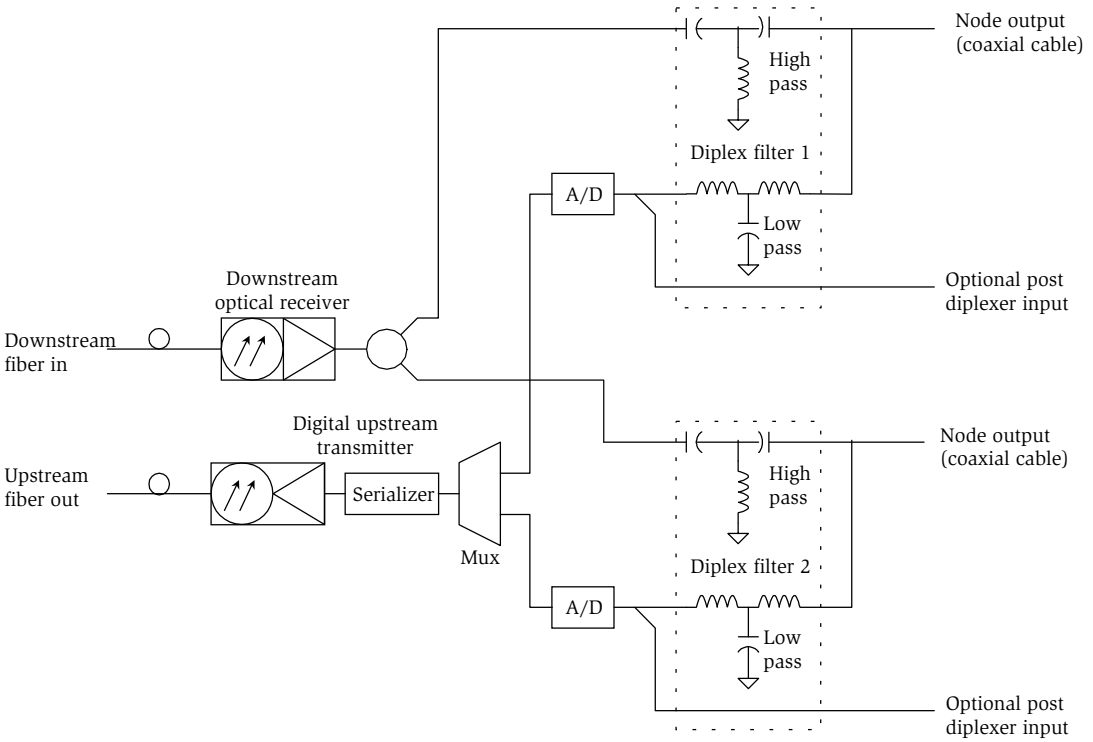


Figure 16.15 Digital return system.

the headend, the digitized signals are converted again to RF in order to allow interface with legacy headend systems.

Figure 16.15 illustrates a digital return system. Compare it with Figure 16.1, which shows the same node with RF return. In the digital return system, the outputs of the two-diplexer low-pass sections are individually digitized in the two *analog-to-digital* (A/D) converters, and then applied to a multiplexer (mux), which alternately passes data from one A/D converter and then from the other. The data is serialized and supplied to a digital transmitter for transport to the headend. At the headend, the data is demultiplexed into two signal streams, which are converted to RF in *digital-to-analog* (D/A) converters. The RF signals can then be supplied to the normal headend upstream receiving equipment, such as DOCSIS CMTSs for data.

As explained in Chapter 6, in order to avoid sampling artifacts, it is necessary to sample data at more than twice the highest frequency to be sampled. If the high end of the return band is 42 MHz, then the sampling rate must be above 84 MHz. A convenient sampling frequency is 100 megasamples per second (Ms/s). Commonly, a 10-bit digital system is used. This means that the data rate from each of

the two A/D converters is 1 Gb/s (100 million samples each second, times 10 bits per sample) before any control overhead is added. Two A/D converters multiplexed together yield a need for a data rate of 2 Gb/s before adding overhead. This fits nicely into an OC-48 transmitter, which runs at a wire rate (the actual data rate) of 2.488 Gb/s. Transmitters and receivers for this data rate are commonly available.

There are several advantages to using digital return transmission. The first is cost. A digital transmitter is lower in cost than is an analog transmitter, because power levels can generally be somewhat lower and because the specifications on a digital transmitter are much looser. Furthermore, by comparing Figures 16.1 and 16.15, you see that only one return transmitter is required, rather than two. Of course, somewhat offsetting the cost reductions is the need for two high-speed A/D converters and the other digital processing circuitry. But you do realize net savings in many cases, at least where you need more than one return path.

Digital returns can be particularly advantageous in longer-distance paths, because the NPR curve does not degrade with distance, at least not until you hit the well-known wall, beyond which digital signals just don't work at all. Yet another advantage is gain stability, since the signal received at the headend is not dependent on the optical signal level received. However, there is inevitable delay in the process of packetizing the signal for transmission, and this delay must be controlled to prevent trouble with some return systems, such as that used in DOCSIS.

Figure 16.16 shows the NPR curve for a 10-bit digital return system on the same scale as is used for the analog return systems shown earlier. The temperature variation is a reflection of the inevitable shift in operating point of the A/D converter and perhaps a function of RF components preceding the A/D converter. It is not a function of shifts in the laser operating point, as is the case in analog transmission. The familiar precipitous drop on the right side is due to saturation of the A/D converter rather than to clipping in the laser.

16.8.3 Return Path Combining at the Headend

In order to save the cost of multiple RF receivers in the headend, it is a frequent practice to combine several reverse path signals into one. This can work if adequate total data capacity (called "bandwidth" by data communications engineers) is available to service multiple nodes as one. This combining is usually done electrically so that there is one optical receiver for each node. RF combining (in the headend, after the optical receiver) can be done satisfactorily as long as adequate carrier-to-noise ratio is maintained, considering both thermal noise and ingress. The transfer gain from the electrical reverse input to the node, to the point of combining in the headend, must be the same for all return paths that are combined. Otherwise, the return operating level of at least one of the nodes will be incorrect.

However, some operators like to do optical combining to further save the cost of optical receivers. Optical combining must be done carefully, if at all, because if the lasers operate at close to the same wavelength then interference

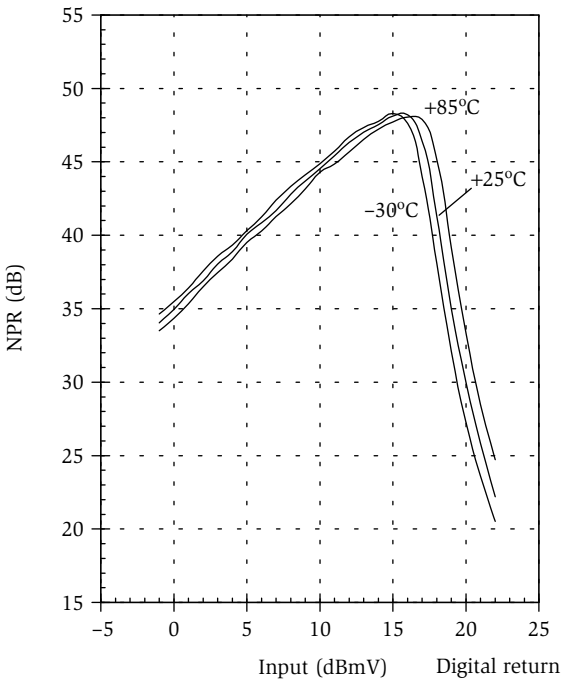


Figure 16.16 NPR curve for a 10-bit digital return path.

(beats) between the lasers will develop in the receiver. These can increase the noise level much more than would be expected from the addition of incoming noise from each transmitter. Also, the received power from each return laser must be the same. Optical combining is not a recommended practice.

Since DFB lasers have relatively narrow ranges of operational wavelengths, it may be possible to combine them. You must ensure that the wavelengths will stay sufficiently separate over life, temperature, and modulation. Because F–P lasers have a less well-defined wavelength operating range, optical combining when F–P lasers are being used is particularly dangerous. Optical combining may work when it is first implemented, but if at a later date the two lasers drift to the same wavelength, the link could fail.

16.9 Spurious Signals in the Return Path

Of paramount concern to operators of two-way plant is noise induced into the return path. Anecdotal and numerical evidence point to the house as being the prime contributor of noise, with the drop contributing the next largest amount.

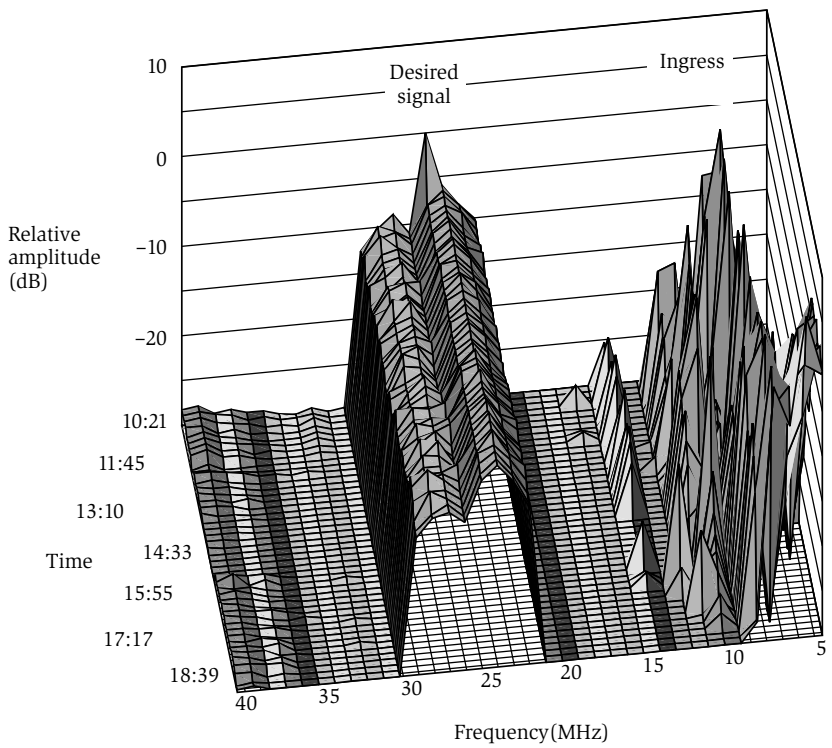


Figure 16.17 Interference versus frequency and time.

Most practitioners contend that the hard-line plant is not responsible for many noise problems, with the exception of common path distortion. A few practitioners have reported noise problems in the hard-line plant, however.

Considerable investigation has gone into this problem of noise on the return plant. Figure 16.17[7] illustrates a fairly typical result. This is a three-dimensional plot of the frequency and time of interference in the return band. Frequency is plotted along the bottom of the chart, progressing from right to left. Along the left edge is plotted time, over a period of about nine hours. The vertical axis represents the relative level of signal observed. This data was taken with a special-purpose receiver, which can sample signals sequentially at a number of frequencies. The data is stored on a computer for later analysis. Other investigators have used a spectrum analyzer interfaced to a computer. A third method for obtaining the data is to capture the return spectrum on a high-speed digital oscilloscope and do a fast Fourier transform (FFT) on the resultant waveform. (An FFT is a mathematical way of computing the frequency spectrum of a complex waveform.)

The figure shows considerable interference at lower frequencies. The frequency spectrum below about 7 MHz is relatively invisible in the upper right-hand corner of the box, but shows less occupancy (by interfering signals) in the morning than in the afternoon. This is consistent with typical off-air usage patterns on these frequencies, which exhibit better propagation characteristics during the later parts of the day and at night. The large energy content between about 22 and 30 MHz is an intentional return signal. The other energy content represents ingress. Above about 15 MHz, this node was essentially ingress free. Below this frequency, interference was substantial and had the distinct probability to cause serious interference with communications.

That the spectrum below about 15 or 20 MHz is relatively "dirty" and the higher spectrum tends to be much cleaner is a fairly typical result. Traditional services, such as status monitoring and return signals from set top terminals using so-called RF-IPPV technology, use very slow and robust signaling methods, such as FSK and QPSK, at low data rates. These are transmitted in this lower portion of the band, where interference is worse. However, the nature of the transmissions is that if one transmission doesn't get through, the system will try again and nothing is lost. (RF-IPPV is the industry acronym for impulse-pay-per-view, IPPV, services that use an RF return channel as opposed to a telephone channel.)

16.9.1 Plant Unavailability Analysis Based on Undesired Signals

Measurement of the quality of the upstream signal path is complicated by the absence of a spectrum of steady-state known signals and by the varying nature and presence of the undesired signals.

Generally, the dominant interfering signals are common path distortion (CPD), discrete ingress signals, and electrical transients. The first two result in signals that may be present for relatively long periods, but occupy narrow frequency bands, whereas transients are present for a very short duty cycle, but may have broad spectral content.

It is important to be able to measure these two classes of signals separately since their effects and the required countermeasures are different:

- Because discrete signal interference varies with time and frequency, the most effective countermeasure is often the ability of a service to shift the upstream frequencies of field transmitters as required to avoid these signals. In order to do that, the headend equipment must continually monitor the spectrum from each node and keep a list of available frequency bands.
- Because electrical transients are typically very brief in duration, the most effective countermeasure is often some combination of interleaving (to prevent one transient burst from affecting more bits than can be corrected by the error-correction capability of the FEC), combined with adequate FEC techniques to correct isolated errors.

The following two measurement techniques are offered as examples of emerging industry practice in this area. The first measures the entire spectrum and range of possible operating levels, but includes only discrete signals, whereas the second measures multiple types of interference, but on only a single channel at a single defined operating level.

16.9.2 Discrete Interfering Signal Probability (DISP)

An NCTA recommended practice for quantification of discrete interfering signals is discrete interfering signal probability (DISP).[8] This procedure was developed by Large and Bullinger as a result of extensive field testing.[9]

Continuous spectrum analyzer measurements are taken over the entire upstream spectrum. The operating parameters are set up to reject transients so that the measurements reflect only discrete signals. The level reading at each frequency on each sweep is downloaded to an attached computer for analysis. Each reading is compared with several predefined thresholds to create a three-dimensional matrix whose axes are frequency, time, and level, and whose values are the probability that a signal is present at a given frequency and time whose amplitude exceeds a given threshold.

From this matrix, postprocessing is used to derive overall discrete signal/CPD performance of the system as a function of threshold, system availability as a function of frequency and operating level, channel availability as a function of time and operating level, and other data as required.

16.9.3 Plant Unavailability Analysis Based on Threshold Boundaries

Studies have been made of the availability of return plant using a CW tester developed at Cable Television Laboratories (CableLabs).[10] The technique uses a CW carrier in the return band and analyzes the characteristics of the signal received at the headend to predict the error rate of a signal at that frequency and with that bandwidth. From that data, we can predict the percent availability of a channel. Whereas the DISP measurement was based on interference amplitude measurement, the CableLabs technique analyzes the errors in location of a point in the data constellation (see Chapter 4 for an explanation of the data constellation).

Figure 16.18 illustrates the CW tester block diagram and the resulting constellation. A CW carrier is added somewhere in the node to be studied. At the headend, the carrier is up-converted to the normal TV IF of 44 MHz and passed through a SAW filter. After AGC (not shown), the signal is supplied to a phase locked loop, which generates quadrature demodulating carriers (see Chapter 4).

The outputs of the demodulator are supplied to a digital-sampling oscilloscope and to a window comparator that determines when the signal is outside the accurate decoding window. The accurate decoding window is shown in the figure; as shown, it is the decoding window for 16-QAM. While the constellation

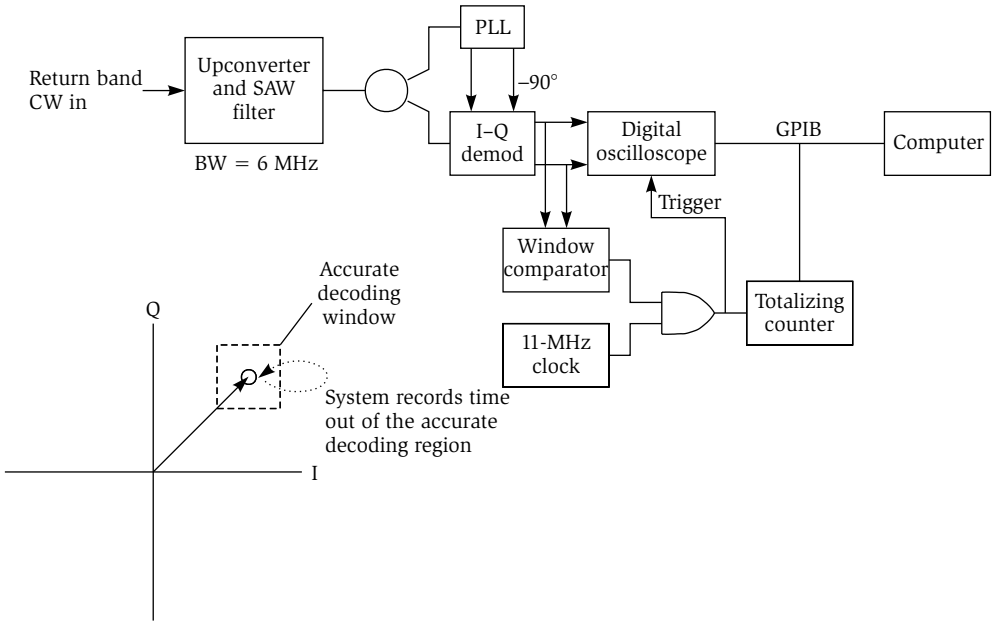


Figure 16.18 CW tester block diagram and constellation.

point is in the window, the correct state is decoded. While it is outside the window, a totalizing counter counts pulses from an 11-MHz clock. The output of the totalizing counter is accumulated by a computer, along with output data from the oscilloscope.

The constellation diagram at the bottom of the figure illustrates the data point crossing the threshold to the region occupied by another state. Assume for a moment that the study concerns use of 16-QAM modulation on the return carrier. The box defining the threshold region, in which the signal will be correctly demodulated, is shown. If a disturbance of any kind causes the carrier to go outside the box, then the totalizing counter is started, and it measures the total time the signal is out of the box. From this information, we obtain the symbol error rate (SER) equal to the proportion of time out of the accurate decoding region. If we assume that the channel is available when the SER is below some level (10^{-5} in the example shown), then we can chart the percent of unavailability of the channel.

Notice that the parameter measured is called the symbol error rate rather than the bit error rate. This is because it is the symbol that is subject to error, and typically each symbol comprises two or more bits. Chapter 4 explains the differences between bits and symbols in the context of data transmission.

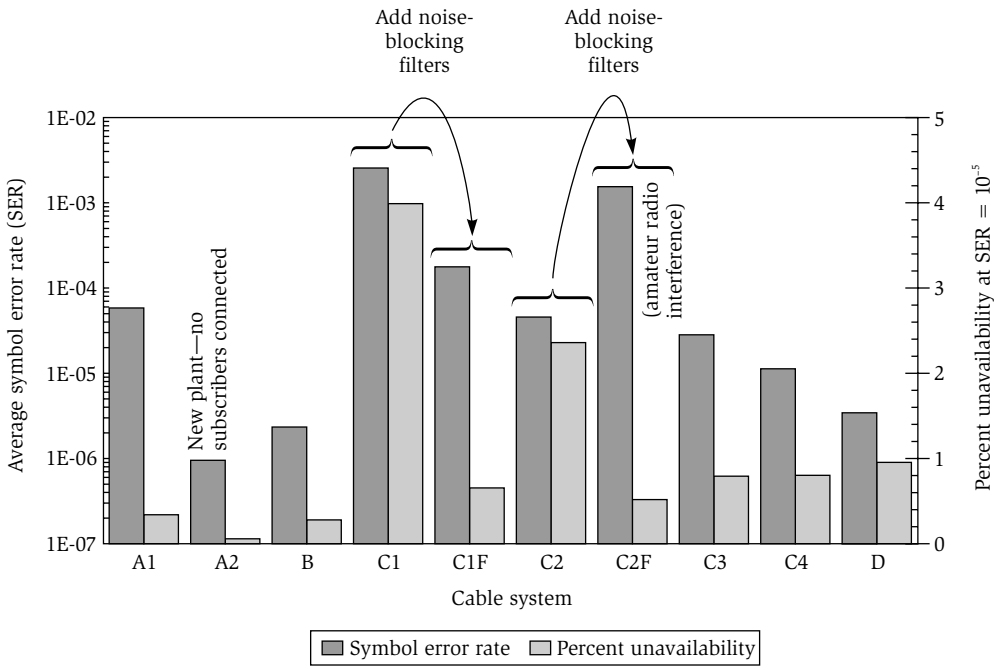


Figure 16.19 SER and percent unavailability of selected nodes.

Figure 16.19 illustrates the symbol error rate and consequent percentage of unavailability of nodes in several cable systems based on one set of tests. Each was observed for about one week. To the left is a logarithmic scale showing the symbol error rate, and to the right is a scale showing the percent of unavailability. Note node A2, which was a new node with no subscribers. Its good performance is evidence of the observation made by others — that drops and house wiring tend to be responsible for many return path problems.

Nodes C1 and C2 were studied with and without noise-blocking filters. Noise-blocking filters are high-pass filters placed on taps to preclude in-home–generated noise from reaching the return path. In node C2, interference developed from an amateur radio operator during the testing; nonetheless, the percent of unavailability dropped significantly when the filters were added. The symbol error rate and unavailability improved dramatically in node C1 when the filters were added. This observation again lends credence to the hypothesis that the home is responsible for most of the problems in return path operation. It also attests to the efficacy of noise-blocking filters (covered in more detail later in the chapter).

16.10 Characteristics of the Composite Reverse Signal[11]

A sinusoid (CW signal) has a well-known peak-to-average ratio given by $20 \log \sqrt{2} = 3$ dB. A composite signal composed of n carriers of equal level and randomly phased, has a total average power n times the power of one carrier, or $10 \log n$ dB greater than one carrier. The peaks add on a voltage basis (because at some time they will all add in phase), so the peak power is $10 \log (2n^2)$ dB greater than the average power of a single carrier. As more carriers are added, the peak power compared with the RMS power increases though the peak is reached for progressively shorter times.

This peak-to-average ratio argues again for the characterization of return lasers using NPR techniques rather than discrete carriers. When a number of carriers are combined, as for transmission on a return optical transmitter, the composite signal is noiselike in its amplitude and frequency characteristics. A noiselike signal has a well-defined and measurable RMS power, but it doesn't have a distinct peak value. Very occasionally, a high peak comes along, which will cause the laser (and electronic amplifiers) to clip.

As we add more carriers to the signals applied to a laser, the more noiselike the signal becomes in terms of peak-to-average ratio. Figure 16.20 illustrates the addition of carriers. Plotted on the x axis is the probability (percentage of time)

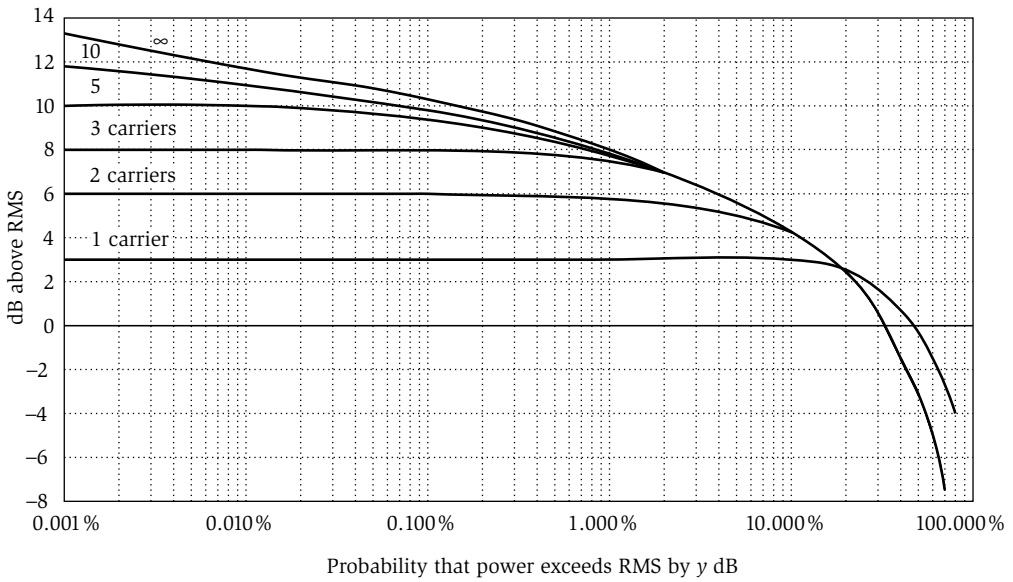


Figure 16.20 Peak-to-average ratio of multiple CW carriers.

that the waveform will exceed the RMS value by the number of decibels shown on the *y* axis. For example, look at the *x*-axis value of 10%. Following it up to the 3-carrier line shows that a signal consisting of three (sinusoidal) carriers exceeds its RMS value by just over 4 dB for 10% of the time. If the total signal comprises more than three signals, it will exceed its RMS value by at least 4.5 dB for 10% of the time.

As the graphs level off (toward the left side of the figure), it is proper to speak of the *y* axis as representing the peak-to-average ratio. For three carriers, the peak-to-average ratio is about 8 dB, and for five carriers, it is about 10 dB. We cannot see the peak-to-average ratio for a larger number of carriers because it occurs for a smaller fraction of the time than 0.001%, the leftmost value on the graph.

As the number of carriers is increased, the curve approaches that of random noise, which has a distribution equal to the infinite curve. This is a confirmation of the central limit theorem from statistics. Stated in engineering terms, the central limit theorem may be expressed as follows:

> The sum of independent variables X_j will be normal for $n \rightarrow \infty$, even if the X_j each have a different distribution, provided that . . . the variance of any one term is negligible compared to the variance of the sum. If this is not so, the term whose variance is not negligible will predominate in the sum and its distribution will still be apparent in that of the sum.[12]

This demonstrates one of the shortcomings of evaluating the return laser with four CW carriers, as described earlier. The peak-to-average ratio of four carriers is around 9 dB, much less than the peak-to-average ratio of random noise or, equivalently, a large number of independent signals sharing the return spectrum. A composite signal consisting of 10 carriers will have a peak-to-average ratio of about 13 dB, within 2 dB of the peak- (probability 10^{-8}) to-average ratio of random noise, so this signal may be acceptable for testing purposes.[13] However, in terms of peak-to-average ratio, random noise remains the most appropriate test signal.

16.11 The Reaction of Active Components to Signal Characteristics

The "peaky" nature of the signals has caused a great deal of concern within the engineering community regarding the reaction of laser transmitters to the peaks in the signal. At least two differing ways to analyze the problem have been used. These are summarized here.

16.11.1 Effect of Laser Clipping—Frequency Domain View

One can analyze the effect of laser clipping in the frequency domain, as done by West.[14] Figure 16.21(a) illustrates no laser clipping. The laser transfer characteristic

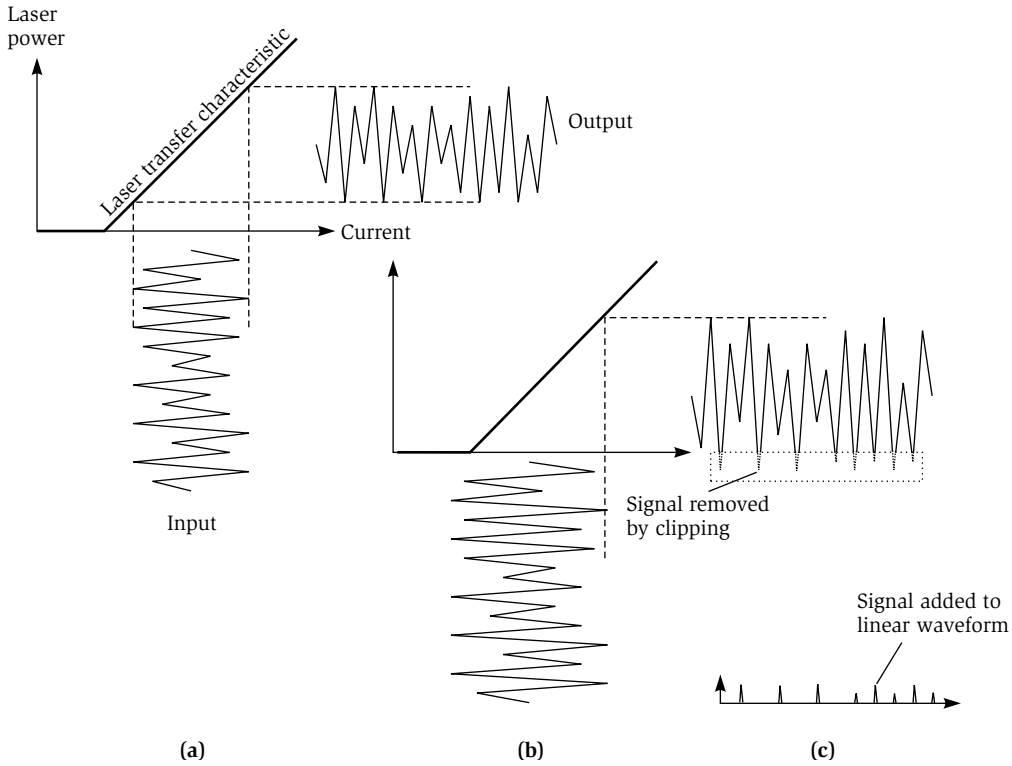


Figure 16.21 Laser in linear region and in mild clipping. (a) No clipping. (b) Clipping. (c) Quasi impulse of clipping signal.

curve shows that, with laser bias current (the x axis), there is no light output until a threshold current is reached, above which the light output power is proportional to the current through the laser. The input signal is shown below the transfer curve, and the output is shown to the right. So long as the input signal remains such that the light output never drops to zero, changes in the output light are proportional to changes in the input current.

As shown in Figure 16.21(b), if the input signal current is such that the light output ceases for a portion of the input signal, then the light output "clips" the signal. The portion of the signal represented by the dotted line is not really transmitted, so the signal is distorted. One way to model this is to assume that the signal was converted to light with no clipping, but then a signal was added, exactly equal in amplitude and opposite in phase, to the portion of the signal clipped. This added signal is shown in Figure 16.21(c). It represents a series of near impulses, which have a frequency spectrum that extends from some very low frequency equal to the lowest beat note

between any elements of the input signal to very high frequencies. Thus, the spectrum of the clipping energy extends from very low frequencies to frequencies generally exceeding the bandwidth of the return path. The result is noise degradation, as represented by the degradation in NPR on the right side of Figure 16.16 and earlier figures.

16.11.2 Effect of Laser Clipping—Time Domain View

Kenny[15] has demonstrated the effect of clipping on a single data signal being carried in the reverse path. The effect is extended to more than one carrier by the central limit theorem. He found that, with a variety of lasers, the effect of clipping is to compress the signals, which compromises bit error rate, as shown by West.

As shown in Figure 16.22, the test setup is similar to that for the NPR test of Figure 16.9. A signal from CW generator G2 is placed in the noise notch, and in one test the noise source, G1, is replaced by a second signal generator. The observation of test results follows a technique developed by Cable Television Laboratories, which is described in Section 16.9.3. A CW carrier may be thought of as "1-QAM" modulation: a single state in a constellation diagram. This is not useful for communicating information, but much can be learned about the communications channel with this test. (For a discussion of levels of QAM modulation, see Chapter 4.)

Generator G1 is adjusted to sufficient amplitude to cause the laser to clip. The "signal" is demodulated by quadrature detectors, which produce in-phase (I) and quadrature (Q) components. (Chapter 4 describes the resultant constellation display in detail.) The phase of the two channels is adjusted to place the recovered "information" in the first quadrant, as shown on the oscilloscope screen. As expected, the spot that should represent the state of the modulation is smeared owing to the noise that fills in the notch, as shown in the description of NPR testing. The smear is circular, indicating phase effects as well as amplitude effects, due to clipping noise falling on the frequency of the test carrier. The location of the smear is toward the origin, indicating a reduction of signal level as expected.

In a second test, the noise source was replaced with a CW generator, whose frequency was chosen such that distortion products did not fall within the passband of the receive filter (the frequency of the test carrier). The laser was forced into clipping by increasing the amplitude of the CW generator that replaced the noise generator, and the resultant constellation spread is as illustrated in Figure 16.22(c). The spot turned into a line along the radial from the origin, again indicating that the laser was clipping the desired signal, but showing the lack of intermodulation products at the test frequency.

When a laser is in clipping, its light is cut off for some short time. During this time, all signals being transmitted by the laser are cut off, not just that largest

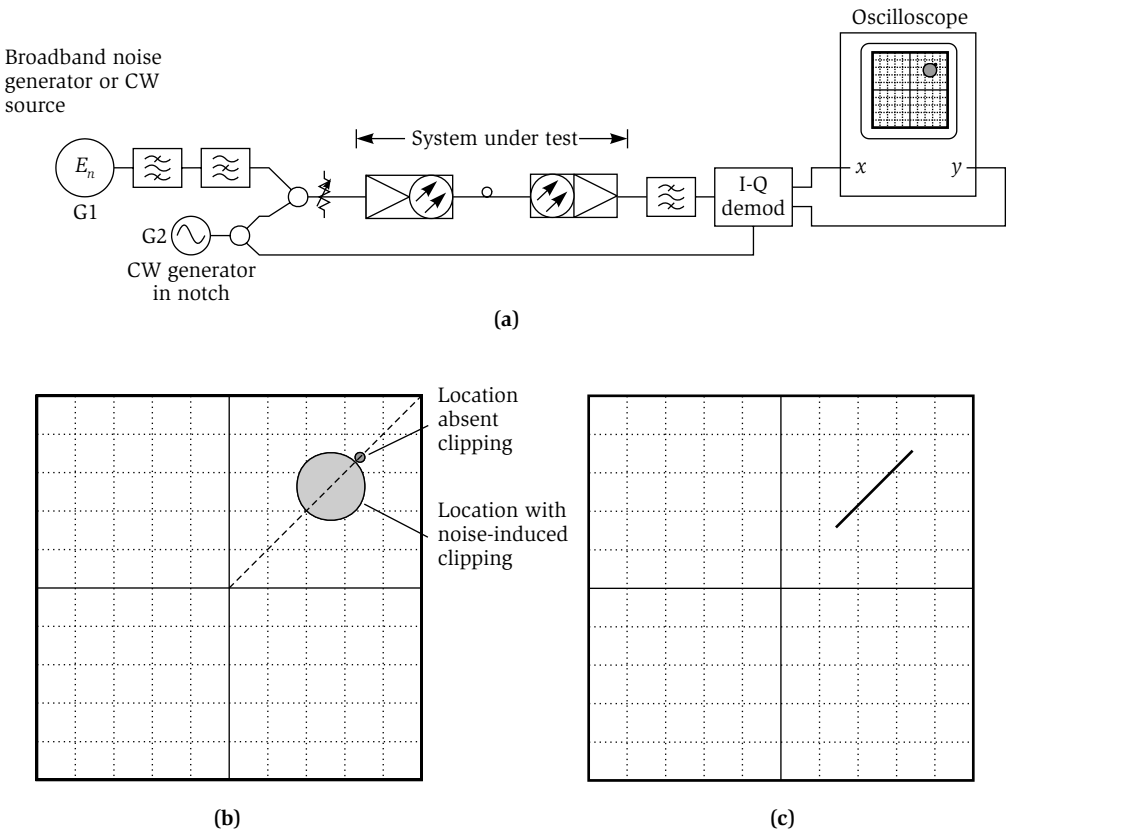


Figure 16.22 Time domain analysis technique. (a) Test configuration. (b) Constellation with noise causing clipping. (c) Constellation with CW signal causing clipping.

signal. Since the signal is cut off for some time, typically for a fraction of a cycle, the amplitude of that signal will be reduced. In the display of Figure 14.22, reduction in amplitude results in the spot on the oscilloscope moving toward the center of the screen, as shown in (c).

16.11.3 Amplifier Characteristics

RF amplifiers tend to react similarly to the peaky nature of the return signals. Typically, the dynamic range in which satisfactory performance is achieved is somewhat better than with return lasers, but many of the same concerns apply. Currently, there is considerable investigation into the ideal characteristics of return amplifiers.

16.12 Common Path Distortion

Common path distortion (CPD) has gotten a lot of attention recently. It manifests itself as a series of beats in the return spectrum, located at multiples of 6 MHz (for North American systems using 6-MHz channel spacing). The distortion is a result of intermodulation of the downstream signals. The cause has usually been found to be mechanical connections at passive components in the common path of the upstream and downstream signals. Examples of such points are connectors, taps, drops, and terminators. These components handle the combined upstream and downstream signals, as opposed to amplifiers and nodes, which separate signals bound in the two directions.

Any mechanical junction is potentially a source of CPD if it handles signals in both directions. If the junction exhibits any nonlinearity, then intermodulation develops. Corrosion on the mating surfaces will typically create such a nonlinearity. Mitigation consists of careful selection of materials on mating surfaces to minimize corrosion and to ensure that the connections are airtight, meaning that they are held in place with so much force that there is a region that is always not in contact with air so that corrosion cannot take place.

This latter criterion is not easy to achieve. Often, because of temperature fluctuations or mechanical vibrations, a point on a contact that is airtight at one time will at another time be exposed to air. Corrosion can develop whenever water (in the air) is present. This mechanism has been called *fretting corrosion*.

There have been reports of components that were assembled with small contacts between, for example, ground and active components. Small contacts with relatively little force holding them together are particularly susceptible to fretting corrosion, especially if inappropriate materials are used. However, the majority of devices using mechanical connections work quite well today: manufacturers have had extensive experience with the problem and are able to control it well.

16.13 Return Path Interference Mitigation Techniques

A number of proposals have been made for noise mitigation on the return path. Some operators who have reasonably new and properly installed plant feel that noise mitigation of any sort is unnecessary. However, it is not possible for the cable operator to maintain control over in-house wiring over time since the subscriber has the legal right to provide his or her own wiring. Traditionally, the availability of poor-quality coaxial cable and connectors in the consumer market, coupled with lack of proper training and installation tools, has resulted in degradation of in-home wiring over time. Some success has been achieved by operators who make available installation kits having high-quality components and, perhaps, instructions for proper installation. Some operators have also worked with local retail establishments to ensure that they carry quality components.

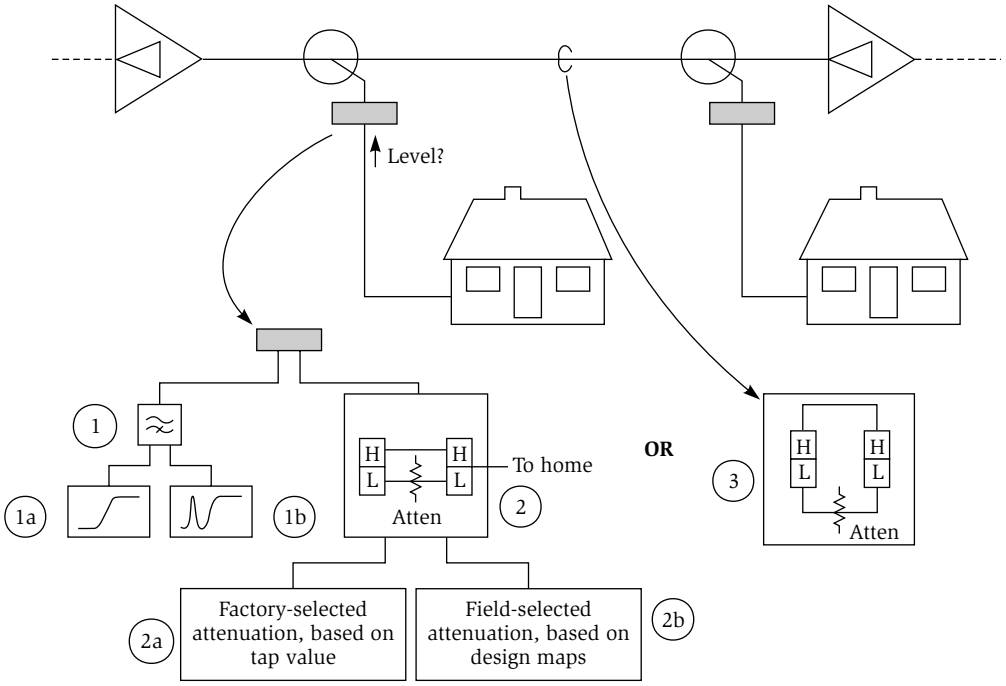


Figure 16.23 Options for interference mitigation.

Figure 16.23 summarizes a number of proposed mitigation techniques, all of which are now being considered. They involve putting something in the hard line between amplifiers or at the drop (or in some cases, at a splitter in the house). Several approaches have been proposed, and they are summarized here.

16.13.1 Option 1: Drop Filters

Drop filters can be used to remove return band energy that might be coming out of the house. Two alternatives, 1a and 1b, are shown in Figure 16.23. Option 1a is a high-pass filter that cuts off below about 54 MHz, the low end of the forward band. It is intended for use at homes not taking any services requiring two-way facilities. With typically 25 to 40 dB of attenuation in the return band, it does not allow return band signal power in the home to reach the hard-line plant.

Filter option 1b is similar to 1a except that a bypass is added to allow signals in a narrow band, often between 10 and 15 MHz, to pass. These are used when the home has an RF-IPPV converter, which must communicate back to the headend on these frequencies. Interference from the home that occurs in the

passband obviously appears in the hard-line plant, but interference coming out of the house at other frequencies is blocked.

Filters are usually placed at the tap though they may be placed at the side of the home in some cases. They are sometimes seen as a stopgap measure because they don't remove the root cause of the problem, and they interfere with provision of modern two-way services in the home. When it becomes the practice to allow the customer to purchase two-way hardware at retail and install it himself or herself (a desirable way to sell data services), the filter can cause particular difficulties. Also, the filter is seen by some as an undesirable expenditure at homes not paying for enhanced two-way services. Countering this last argument is the thought that use of the filters postpones the day that homes not taking two-way services will have to have their internal distribution wiring improved.

Even when two-way services are added to the home, it may be possible to split signals at some point, with a new cable supplying only the device requiring return facilities. The filter can then be placed in the path to all services not requiring two-way facilities.

16.13.2 Option 2: Return Attenuation

A second interference mitigation option is to provide, in the tap or externally, two diplex filters, which separate the upstream frequency band (5–40 MHz) from the downstream band (54 + MHz). An attenuator is inserted into the low side of the diplex filter chain so that the loss is increased in the return direction while not changing the loss in the forward direction. The rationale is that tap values are selected to yield the correct signal level at the highest downstream frequency. Since cable loss is roughly proportional to the square root of the frequency, the loss in the upstream direction is considerably lower than in the downstream direction. Thus, whereas the *downstream received* level is the same at all houses (within a certain range), the *upstream transmit* level is significantly different between homes.

Figures 16.6 and 16.7 illustrate the range of levels that could be expected in a particular node that was analyzed for return level. Generally, as you move downstream from an amplifier, the tap value is reduced to compensate for signal loss at the high end of the downstream band. Since return signals suffer much less loss, the required upstream transmit level is reduced as you move downstream from an amplifier. Since the taps farther from an amplifier require lower transmit level, they also admit more potential interference from the house.

The thought behind adding the attenuator in the return direction is that this will bring taps requiring a low signal level up to requiring a higher signal level. In the process, it will attenuate interference coming out of those homes. It would not deter return services from being installed, as might filters. However, the diplexers will add some group delay.

Option 2a: Factory Selection of Attenuation Based on Tap Value

Within the spectrum of options for the return band attenuators is the possibility that the value of return attenuation is selected at the factory, based on the tap value. This would not place additional burden on field installation personnel, who treat the tap as they would any other tap. The problem with this approach, however, is that the range of transmit levels required at any particular value tap (Figure 16.6) is so great that factory selection of an attenuator value would be very difficult.

Figure 16.24 illustrates an attempt in one real node to add reverse attenuation. In this example, all tap values 23 dB and below had reverse attenuation added such that the total attenuation in the reverse spectrum equaled 23 dB. Dots show the required levels before addition of the reverse attenuation, and + marks show the required levels after. Notice that after adding attenuation, some of the low-value taps now require the highest return levels. The reason is that some low-value taps are placed after a lot of cable attenuation, so they benefit from a lot of added attenuation. Other low-value taps follow a directional coupler so that the loss is primarily flat (frequency insensitive) loss, and the return signal level required after addition of the attenuation is high. This is not a good condition since it could require excessive output level from a transmitter.

Preliminary analysis has raised questions about the efficacy of factory selection of attenuator values. If the amount of attenuation added is low enough to prevent possible excessive level requirement, the benefit derived is minimal.

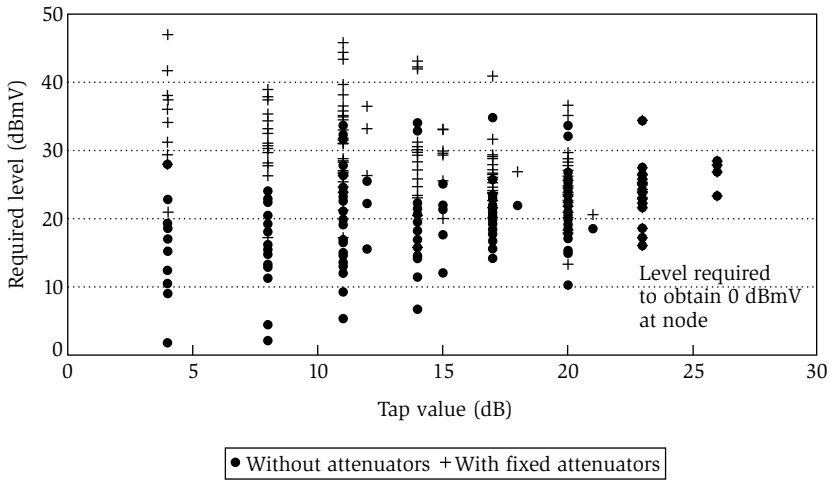


Figure 16.24 Required tap level with and without reverse attenuation.

Option 2b: Field Selection of Attenuation Based on Design Maps

The other option for attenuators in taps is to provide a plug-in attenuator that is selected in the field, based on computed gain back to the previous amplifier. This technique shows considerable promise to improve the ingress problem while permitting return transmitters to be designed with a very narrow return signal range. The problem is that additional burden is placed on field installation personnel, who first have to select the tap based on system design and then have to select the correct attenuator value to plug in. After the tap is installed, it is not possible to inspect the attenuator for the correct value. If the tap value is changed later, the new tap may not be equipped with the optimum attenuator value.

16.13.3 Option 3: Move Diplexers and Attenuator to Midspan

Others have suggested that the diplexer be moved from the tap to approximately midspan between amplifiers, and the attenuation used be selected to make the transmit levels required downstream of the diplexer to be roughly equal to those before the diplexer. This would reduce the number of diplexers required by somewhere around a factor of three or four, compared with putting them in taps. It does necessitate inserting another device into the hard-line cable. Co-locating the diplexers with an equalizer has been suggested, but this does not seem to be optimum.

Preliminary analysis indicates that such placement, combined with optimum attenuation selection in the field, would significantly reduce the signal level range required of transmitters. However, the estimated improvement in ingress is not particularly good, raising questions about the efficacy of the technique.

Notice that the industry has not firmed up its philosophy for applying these ingress countermeasures. Future work may reveal advantages that are not obvious at this time and may result in newer techniques that yield more benefit. Also, it is possible that the industry will learn how to improve performance of problem homes without having to use any of the techniques shown in this section.

16.13.4 Option 4: Frequency Hopping

Many services that use return plant service have provisions for frequency hopping. They can move from one frequency to another if their original frequency develops interference problems that make communications unreliable. This subject is covered in some detail in Chapter 6.

16.13.5 Option 5: Error Correction

Error correction is often added to digital datastreams to allow correction of any bits that are corrupted in transmission. The effect is to sharpen the curve of error rate versus carrier-to-noise (or interference) ratio. As noise increases, the

recovered signal retains its quality longer, but at some point, the quality will collapse rapidly with worsening carrier-to-noise ratio. Chapter 3 describes error correction as it is being applied in the cable industry.

16.14 Examples of Signal Apportionment

We showed earlier that an optimum operating level for a return path laser exists. It is desirable to set the actual operating level as close to this as possible without going over the limit where clipping gets to be a significant issue. The operator must take into account temperature effects, errors in level setting, and laser loading due to ingress. He or she must also account for future services that may be added; it is not good to go back and readjust operating levels to accommodate a new service. After the operating point has been determined by selecting an appropriate operating point on the NPR curve, it remains to apportion power among multiple services intelligently. This section gives two examples of how an optical transmitter could be operated under different conditions.

16.14.1 Few Return Services, F–P Laser

The first example illustrates the apportionment of signal power for a fairly simple node having only a few return path services and using a Fabry–Perot (F–P) laser. The laser is the one illustrated in figure 16.12. The chosen operating point is a total signal power into the transmitter of +15 dBmV, where the laser exhibits about a 42-dB NPR at room temperature. In practice, you may want to operate slightly lower on the curve to allow more room for error, but we take this point for illustration.

The services assumed include the following:

1. A status-monitoring carrier using FSK (or it could be BPSK) modulation at a center frequency of 7 MHz, with an occupied bandwidth of 300 kHz (total, not plus and minus)
2. An STT return carrier according to SCTE 55–1 at 13 MHz, with an occupied bandwidth of 192 kHz, QPSK modulation
3. A DOCSIS 1.1 RF return using 16-QAM, centered at 21 MHz, with an occupied bandwidth of 800 kHz

A significant omission from this list, done to keep the example simple, is an allocation for future expansion. It is essential that all future services be allocated the amount of power they will need when added. Failure to do so can result in extreme laser clipping later, and possibly a situation where a node that previously worked no longer does. Also, we have not allowed for ingress. It is advisable to allow some power for ingress.

One technique used to allocate signal level is the so-called *constant power density* method. In this method, each signal is set at the same amplitude per unit of occupied bandwidth. That is, if one signal is twice as wide as another, it will be set at twice the amplitude of the narrower signal.

The signal levels are set such that the total signal power is equal to the total at the operating point of the NPR curve, for example, +15 dBmV in Figure 16.12. The signals must be added on a power basis. If the levels of the three signals are L1, L2, and L3, in dBmV, then the power must add to the target power, in this case +15 dBmV:

$$\begin{aligned}\text{total power} &= 10\log(10^{L1/10} + 10^{L2/10} + 10^{L3/10}) \\ &= +15\,\text{dBmV}\end{aligned}$$

Note that, though signal level is expressed in dBmV, as is customary in the cable television industry, we are actually manipulating signal power, expressed in dBmV in a 75-ohm system. For computation, we must express this level in terms of power, because it is necessary to add the drive power contributed by several signals. One cannot add power when that power is expressed in decibels. In the inner expressions — taking the power level, dividing by 10, and raising 10 to the power of the result — we are taking the antilog of the level expressed in dBmV. Thus, we have legitimate power units, though they are not commonly used units such as milliwatts.

In the apportionment philosophy shown here, the constant power density method is modified to give more power to services requiring a higher carrier-to-noise ratio. See Chapter 4 for a discussion of the carrier-to-noise ratio (C/N) required to meet a certain bit error rate (BER). The power ratios for different modulation formats are based on the distances between state boundaries in the constellation diagrams, with the average level of the signals held constant. In order to exhibit the same BER, a QPSK signal must have a 3-dB-higher C/N than that needed by a BPSK signal. Thus, in the example, the QPSK signal is allowed 3-dB (times 2) more level per hertz than is the BPSK signal. Similarly, the 16-QAM signal needs approximately 7.1-dB (times 5.13) better C/N than does a BPSK signal to yield the same BER. ($10\log(5.13) = 7.1$ dB.) Thus, we shall allocate twice the power density to a QPSK signal than we allocate to a BPSK (or FM, since we are treating the two the same, for better or for worse). We shall allocate 5.13 times the power density to a 16-QAM signal as to a BPSK signal.

It is possible to construct a table that will allow assignment of levels to each of the signals. We can express the power of each signal as follows. Let p represent the power per hertz in the BPSK signal. Table 16.1 illustrates how power is added. We multiply the power per hertz, p, by the occupied bandwidth of the service. The QPSK STT return is assigned twice the power because it is twice ($10\log 2 = 3$ dB) as sensitive to noise as are the signals of the BPSK

Table 16.1 Power Assigned to Each Service, Node with F–P Laser

Service	*BW (kHz)*	*Total Power*
1. Status Mon.	300	$300p$
2. STT return	192	$2 \times 192p$
3. DOCSIS 1.1	800	$5.13 \times 800p$
Total power		$4788p$

services, and so on. With this weighting factor, the total power, shown at the bottom of the chart, is $4{,}788p$. Equating this to the total power to be distributed among the services, +15 dBmV, we obtain

$$\begin{aligned} 4788p &= 10^{15/10} \\ p &= 6.6 \times 10^{-3} \text{power units per hertz} \end{aligned} \tag{16.1}$$

We are working legitimately in power, but the units are not common units, so we just call them power units. Since we will convert back to dBmV consistently, we can get away with any arbitrary power units. We simply convert to power in order to add several signals. This may be converted back to dBmV:

$$\begin{aligned} \text{power density} &= 10\log(6.6 \times 10^{-3}) \\ &= -21.8\,\text{dBmV per kHz} \end{aligned} \tag{16.2}$$

We can then determine the total power of each BPSK signal by adding 10 log(bandwidth) to the signal level in dBmV. For the status-monitoring signal (300 kHz wide), this gives a total power of

$$P_{sm} = -21.8 + 10\log(300) = 2.97\,\text{dBmV}$$

(Note that we took bandwidth in kilohertz, to be consistent.)

We can estimate C/N as follows. From Figure 16.12, the NPR at the selected operating point is 42 dB. This means that the noise per hertz is 42 dB below the signal level per hertz (this is the definition of NPR). The total signal power at the selected operating point is +15 dBmV, and this signal was assumed to be spread over 35 MHz (5–40 MHz). Thus, the signal power per hertz is

$$+15 - 10\log(35{,}000{,}000) = -60.44\,\text{dBmV/Hz}$$

Since the noise level is 42 dB lower yet, it is −102.44 dBmV/Hz.

The status-monitoring signal is assumed to be 300 kHz wide, so the noise level is $10 \log (300{,}000) = 54.77$ dB higher than -102.44 dBmV/Hz; that is, the noise level in the passband of the status-monitoring signal is $-102.44 + 54.77 = -47.67$ dBmV. This compares with the previously computed signal level of 2.97 dBmV, so

$$\mathrm{C/N} = 2.97 - (-47.67) = 50.64 \text{ dB} \tag{16.3}$$

This is a bit optimistic, because we assumed that the receive filter for the status-monitoring system has zero excess bandwidth. This is not realistic, so we will lose some C/N because the filter is wide enough to admit somewhat more noise than we assumed. In this case, the C/N is so far above what we need that we have nothing to worry about. A more precise computation would have to take this excess bandwidth into account, though.

Thus, we have so far established that the total amplitude (when the entire signal is measured) of the status-monitoring signal is +2.97 dBmV at the input to the return transmitter and that the carrier-to-noise ratio is 50.6 dB. It remains to determine what we would see on a spectrum analyzer if we looked at this signal. In order to determine this, we must make some assumptions concerning the way the analyzer works and how it is set. For the purpose of illustration, we assume that the analyzer is set to a resolution bandwidth of 100 kHz, narrower than any of the real signals being observed. We further assume that the noise bandwidth of the analyzer is 1.2 times the resolution bandwidth, or 120 kHz. The ratio of resolution bandwidth to noise bandwidth depends on the types of filters used in the spectrum analyzer. We will not use a detector correction factor, though the analyzer manufacturer may have one that should be taken into account; each analyzer model can be different in this respect. From the foregoing, we have the level of the status-monitoring signals as 2.97 dBmV measured in a bandwidth of 300 kHz, so if it is observed in a bandwidth of 120 kHz, the level we should see is given by

$$\begin{aligned} \text{level}_{\text{analyzer}} &= 2.97 - 10 \log \frac{300}{120} \\ &= -1 \text{ dBmV} \end{aligned} \tag{16.4}$$

In a similar manner, we can determine the operating point of the STT return QPSK and the DOCSIS 16-QAM signals. When we compute the amplitude of the STT return signal, we must recall that it is twice what we compute for the BPSK signals, as a result of adding 3 dB (that is, multiplying by 2) to the level of the BPSK signals. The amplitude of the 16-QAM signal is multiplied by 5.13 to account for its greater susceptibility to noise. The results are shown in Figure 16.25. For each of the three signals, the higher, thicker line indicates the true amplitude of the signal when measured over at least its occupied bandwidth. The width of the line indicates the occupied bandwidth. Beneath each signal is

a thinner line that represents what a spectrum analyzer with the specified characteristics would read. In order to correct the analyzer reading for the actual signal level, we would have to add to the reading 10 log (ratio of occupied bandwidth to measuring bandwidth).

Further, note that the analyzer reading for the 16-QAM signal is substantially below the actual signal level, because the 16-QAM signal is so much wider than the analyzer's measurement bandwidth. The difference is less dramatic for the QPSK signal, because it is not as wide as is the 16-QAM signal. Also, note that the difference between the analyzer readings for the QPSK and BPSK signals is 3 dB, the level by which they differ on a per-hertz basis.

Figure 16.25 is not to be taken as a spectrum analyzer display: The lines with the labels next to them show the frequency and total amplitude of the signal. The lines below show what a spectrum analyzer would display *under one set of conditions*. Under other conditions, the spectrum analyzer will display a different level.

Also note that the example did not include an allowance for ingress, nor did it include an allowance for errors in level of the various return signals. As shown in Section 16.6.2, the signal levels at the return laser transmitter are controlled from the headend. Inevitably there will be some error in the control of the levels. Finally, allowance was not made for future services; you would be wise to allow for such future expansion of offerings.

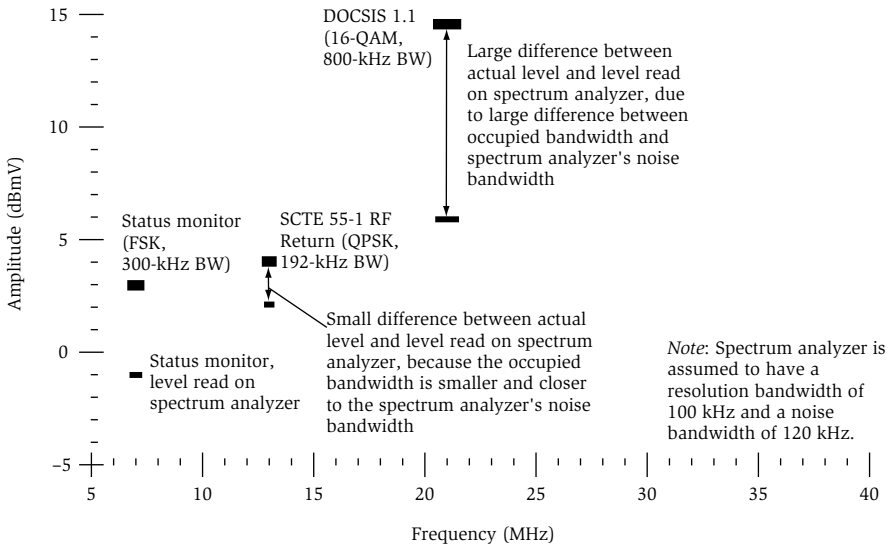

15
10
5
0
–5
Amplitude (dBmV)
DOCSIS 1.1 (16-QAM, 800-kHz BW)
Large difference between actual level and level read on spectrum analyzer, due to large difference between occupied bandwidth and spectrum analyzer's noise bandwidth
Status monitor (FSK, 300-kHz BW)
SCTE 55-1 RF Return (QPSK, 192-kHz BW)
Small difference between actual level and level read on spectrum analyzer, because the occupied bandwidth is smaller and closer to the spectrum analyzer's noise bandwidth
Status monitor, level read on spectrum analyzer
Note: Spectrum analyzer is assumed to have a resolution bandwidth of 100 kHz and a noise bandwidth of 120 kHz.
5
10
15
20
25
30
35
40
Frequency (MHz)

Figure 16.25 Signal level apportionment, F–P laser, limited services.

16.14.2 Many Return Services, DFB Laser

The next example involves a DFB return laser carrying a greater number of signals. Power is apportioned first to the analog signal, giving it the level required to allow it to operate at the desired C/N. Next, the remaining signal capability of the laser is apportioned to the digital signals using a philosophy of constant power per hertz within a type of modulation. As in the first example, the power per hertz allocated to a signal is offset based on the modulation type, with denser modulation formats receiving a proportionately larger share of the power per hertz, to allow for their greater sensitivity to noise.

For the DFB transmitter, we'll select an input level of 18 dBmV and a corresponding NPR of 47 dB as our operating point.

The following signals were included in this example:

1. A status-monitoring carrier using FSK (or it could be BPSK) modulation at a center frequency of 7 MHz, with an occupied bandwidth of 300 kHz (total, not plus and minus)
2. An STT return carrier according to SCTE 55–1 at 13 MHz, with an occupied bandwidth of 192 kHz, QPSK modulation
3. A DOCSIS 1.1 RF return using 16-QAM, centered at 21 MHz, with an occupied bandwidth of 800 kHz

These first three signals are identical to those illustrated in the previous example, Figure 16.25. Several additional signals were added:

4. An analog video return signal from the local high school, with picture carrier at 31 MHz. Its amplitude is set to yield a 50-dB carrier-to-noise ratio from the laser path (noise in other portions of the path will reduce the C/N). Remember that if you are going to handle an analog return signal, the transmitter must be specified to have adequate linearity to keep distortion products out of the analog signal.
5. A QPSK switched circuit telephony signal with an occupied bandwidth of 1.8 MHz.

The digital signals "get" their allocation of power from the power left after the analog signal receives enough power allocation to yield the desired carrier-to-noise ratio.

Figure 16.26 summarizes the level of all signals, using the procedure already outlined. The telephony QPSK signal is the widest of the digital signals, so the difference between its actual level and the level indicated on the spectrum analyzer is the largest difference. The two QPSK signals have the same power density, which is proportional to what is read on the spectrum analyzer as long as its noise bandwidth is less than the occupied bandwidth of the signal.

In this case, the actual level of some of the digital signals is higher than the level of the analog signal. This is simply a result of the way we allocated power

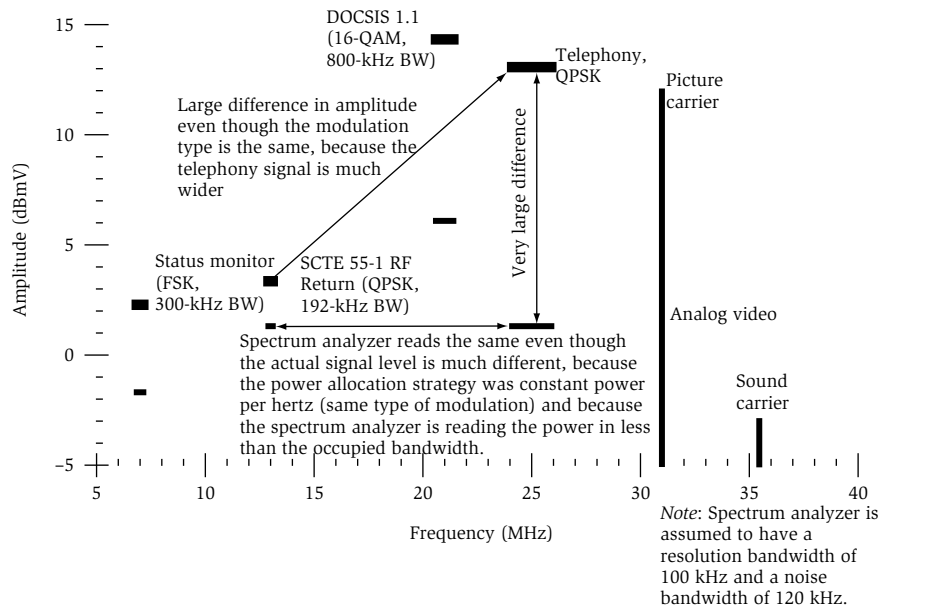


Figure 16.26 Signal apportionment, DFB Laser, many services.

for the analog signal. We set its level to yield a C/N of 50 dB and then subtracted that level from the total +18-dBmV signal level. The remaining signal level happened to be enough that we could allocate more to the digital signals needing the most. Of course, you could say that this link is overdesigned, since the C/N provided is much higher than needed.

16.15 Summary

The chapter has dealt with a number of items of interest in the operation of return HFC plant. It began with a description of the components unique to a two-way plant. We showed key characteristics of those components. Level setting in the return plant received a lot of attention because it is crucial to operation of reliable return plant, just as it is to proper operation of downstream plant. We examined lasers used for return services and gave practical methods of evaluating the lasers. We discussed important characteristics of return signals and their interaction with the laser transmitters.

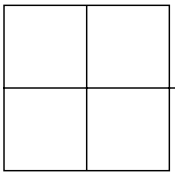
Noise in the return plant is important and received considerable attention, along with proposed mitigation methods. Finally, we gave examples of ways to

set up the levels of various return services and showed how to estimate key performance parameters.

The next chapter deals with system architecture. Methods of plant design are described that optimize performance using the principles shown in this chapter.

Endnotes

1. Data courtesy of H. Carnes, R. Oberloh, and team. Antec, Norcross, GA.
2. National Cable Television Association, Upstream Transport Issues, *NCTA Recommended Practices for Measurements on Cable Television Systems*. The "Upstream" portion was added to the *Recommended Practices* in late 1997. The document is available from the NCTA, Washington, DC.
3. John Kenny, Characterization of Return Path Optical Transmitters for Enhanced Digitally Modulated Carrier Transmission Performance, *Technical Papers of the NCTA Annual Convention*, NCTA, Washington, DC, 1997.
4. Lamar West, Analysis of Reverse Path Laser Loading in HFC Networks. *HFC'97 Conference*, Phoenix, AZ, September 1997. SCTE, Exton, PA.
5. H. Blauvelt et al., Return Path Lasers for High Capacity Hybrid Fiber Coax Networks, *1995 NCTA Technical Papers*. NCTA, Washington, D.C.
6. Donald Raskin and Dean Stoneback, *Broadband Return Systems for Hybrid Fiber/Coax Cable TV Networks*. Upper Saddle River, N.J., Prentice-Hall, 1998. See especially Chapter 9.
7. Data courtesy of D. Junghans, of Arris Interactive, and T. Mitchell, formerly of Arris Interactive. Private communication.
8. Supplement on Upstream Transport Issues, *NCTA Recommended Practices for Measurements on Cable Television Systems*, 2nd ed. NCTA, Washington, DC, October 1997, pp. 57–63.
9. D. Large and R. Bullinger, A Proposed Method for Quantifying Upstream Ingressing Carriers, *1997 NCTA Technical Papers*. NCTA, Washington, DC.
10. Richard S. Prodan et al., Results of Return Plant Testing, *1997 NCTA Technical Papers*. NCTA, Washington, DC.
11. This section is adapted from the *NCTA Recommended Practices for Measurements on Cable Television Systems*.
12. Petr Beckmann, *Probability in Communication Engineering*. New York: Harcourt, Brace & World, 1967, p. 107.
13. National Cable Television Association, op. cit.
14. Lamar West, op. cit.
15. John Kenny, Characterizing Return Path Transmitters. *CED*, May 1997, p. 26ff.



Chapter 17

Architectural Requirements and Techniques

17.1 Introduction

The architecture of a cable system is a description of how the major elements — signal acquisition, signal processing, signal transport, and subscriber terminal equipment — are interconnected. The first step to selecting or analyzing a specific network architecture is to analyze the services to be delivered and their requirements. In a dynamic telecommunications market, however, service penetration and customer usage will grow and new services will emerge. Therefore, the second requirement of the selected network configuration is that its initial cost be consistent with expected revenues but that it be capable of scaling economically to meet market opportunities and customer demands. It is those analyses that will be covered by this chapter.

We discuss first the categories of service-related parameters that are affected by architecture, then the requirements of various specific service types, and, finally, scalability and the interaction of bandwidth assigned to each service with the physical granularity of the network. The emphasis will be on the process for setting standards and on alternative means of meeting those standards rather than on suggesting specific values.

Chapter 18 will deal with specific HFC architectures and how they differ relative to the parameters discussed here. Chapter 19 will cover the same issues for emerging fiber-deep architectures. Chapter 20 will cover the calculation of network reliability and availability, given a specific equipment configuration.

17.2 Performance Parameters

Network performance can be measured by many "yardsticks." Signal quality in the broadband distribution system, including noise, distortion, levels, frequency response, and other related parameters, is extensively discussed in chapters 10–16. Other architecture-affected topics are discussed in this chapter.

17.2.1 Information Capacity

The information capacity of a network that includes an RF FDM layer is a composite of many factors, including the RF bandwidth, the efficiencies of the processes by which raw information is encoded for modulation onto RF carriers, spacial reuse of the network, and efficient sharing among users with varying needs for information transmittal rate. Networks with linear fiber-optic links include another layer, which measures how much modulated RF spectrum can be carried through the optical cables. Not all of these parameters are relevant to some fiber-deep architectures.

Optical Link Capacity

In linear optical links, as discussed in Chapter 12, a composite RF spectrum is used to modulate the amplitude of an optical source at one end of a link, with a matching demodulation at the other end used to recover the modulating spectrum. The maximum bandwidth of the modulating spectrum is limited by the bandwidth and linearity of the modulating devices, by the minimum acceptable C/N of the link, and by the interaction between the optical modulated signal and the fiber medium. The trade-off among bandwidth, signal loading, noise, and distortion is similar to that on coaxial links.

The capacity of an optical link can be increased by simply installing cables containing multiple fibers. In fact, most installed fiber-optic links use, at the least, separate fibers for upstream and downstream communications to nodes. So long as the cables are not bent sharply, the crosstalk between signals on adjacent fibers is extremely low, and cables of less than one-inch diameter can hold approximately 1,000 fibers.

As discussed in Chapter 13, however, it is also possible to send multiple modulated light signals through a shared fiber with a low level of mutual interference, so long as they are on sufficiently separated optical frequencies and the optical power levels are not too high. This is a form of FDM, but since optical wavelengths, rather than frequencies, are usually specified, it is known as *wavelength division multiplexing* (WDM).

Downstream RF Bandwidth

One of the fundamental measures of the broadband, linear distribution network is the range of frequencies used to transport signals toward customers. In residential coaxial distribution networks, the lower limit of downstream frequency range is constrained by the fact that upstream signals are almost always carried below downstream signals in shared cables.* Furthermore, the bands

* One experimental network has been built in which the upstream signals are carried in the 900- to 1,000-MHz band, thereby constraining the upper as well as the lower ends of the downstream spectrum.

must be separated sufficiently that reasonable-cost diplex filters can be deployed in amplifier stations to properly route the different spectra for processing without causing amplifier instability or excessive group delay at the band edges. Additionally, systems that deliver analog television programming to residential customers must include the lowest television channel within the downstream spectrum, whereas those not carrying such programming are sometimes designed with different crossover points between the bands.

In both North and South America, the lowest television channel starts at about 53 or 54 MHz, depending on channelization plan, with the result that the lower downstream frequency limit is typically near 50 MHz (see Chapter 10 for a discussion of amplifier station internal RF circuitry and Chapter 9 for a discussion of typical channelization plans). These are called *low-split* networks. Lower downstream band limits for some nonresidential networks are 150 MHz and 234 MHz, commonly called *midsplit* and *high-split* networks, respectively. For networks requiring more RF bidirectional bandwidth, separate cables and amplifying stations may be used for each direction of transmission with dual coaxial cables, thereby freeing constraints on bandwidth in both directions. Occasionally, dual-cable low-split networks are used for residential networks as a means of increasing both downstream and upstream bandwidths while retaining conventional channel plans.*

At the top end of the spectrum, the bandwidth of coaxial networks is limited primarily by amplifier technology. Chapter 11 contains a discussion of how total RF bandwidth can be balanced against other broadband performance parameters, such as levels, noise, and distortion. The maximum downstream frequency has steadily increased as technology has improved, with 550-, 750-, and 860-MHz amplifiers commercially available. In a few areas, 1,000-MHz systems have been built.

Upstream RF Bandwidth

Historically, few all-coaxial cable television systems were enabled for two-way communication with subscribers, for the very practical reason that large, all-coaxial plants contained so many amplifiers and branches that the combination of noise funneling and ingress made upstream communications problematical. Among those with two-way capabilities, some were used for point-to-point applications, in which the system had few, if any, upstream signal-combining points. Others were used to carry low-information-rate signals, such as data from two-way set-top converters or status monitoring, where only a single transmitter was activated at a time, upstream levels could be high, and rugged frequency shift keying (FSK) modulation was used. In the few systems that

* A number of dual-cable systems were built in the 1980s. The cost of maintaining two networks, however, has led operators to convert most to single cables as they are upgraded.

offered true subscriber interactivity (such as the pioneering Columbus, Ohio, "CUBE" system), electronic switches were employed at field branching points to dynamically reduce the effective upstream combining. Two-way operation is practical in modern HFC systems due primarily to the smaller, electrically separate, broadband distribution subnetworks. Chapter 16 discusses the unique problems with upstream communications.

The low end of the upstream spectrum is limited by the need to separate signals from the harmonics of the powering that is multiplexed with the RF signals. It is also effectively limited by the practical frequency range of the magnetic components used in amplifiers and passive elements, such as power inserters, taps, splitters, and directional couplers. The most common lower frequency limit is 5 MHz.

The upper end of the return spectrum is limited by the need to provide an adequate crossover band between upstream and downstream signals, and also by the need, in residential networks, to avoid feeding high-level signals into the input terminals of television receivers in the common intermediate frequency (IF) band, which is 41–47 MHz. This is discussed in more detail in chapter 24.

Typical upstream band limits in older amplifiers were 5–30 MHz, whereas modern designs are typically 5–40 MHz or even a little higher. For nonsubscriber midsplit and high-split networks, the upper upstream frequency limits are typically 112 MHz and 174 MHz, respectively. These three common spectral usage plans are shown in Figure 17.1.

Bandwidth per Customer

The instantaneous RF bandwidth is one factor fundamentally limiting information capacity when a common set of signals are to be delivered to all customers, which was exclusively the case in the early history of cable television and true for much of the programming on most cable systems today. Increasingly, however, new services require communications paths to individual customers. To the extent that the network can be logically or physically divided into independent transmission systems, we can use the same frequencies to simultaneously deliver differentiated services to subscribers in different network sections. This is known generically as *space division multiplexing* (SDM).

A measure of the network's ability to deliver customized services is the *bandwidth per customer*. In the downstream direction this is simply equal to the instantaneous RF bandwidth devoted to some service divided by the number of homes in the smallest portion of the network that can be fed unique signals multiplied by the penetration of that service among homes passed. In the upstream direction, it is the bandwidth devoted to that service divided by the number of customers sharing that bandwidth at any point along the path. The upstream and downstream per-customer bandwidth requirements may be symmetrical (for instance, for telephone) or highly asymmetrical (for most interactive video services). Since service penetration changes with time, network capability is often measured in terms of achievable bandwidth per home passed rather than bandwidth per customer.

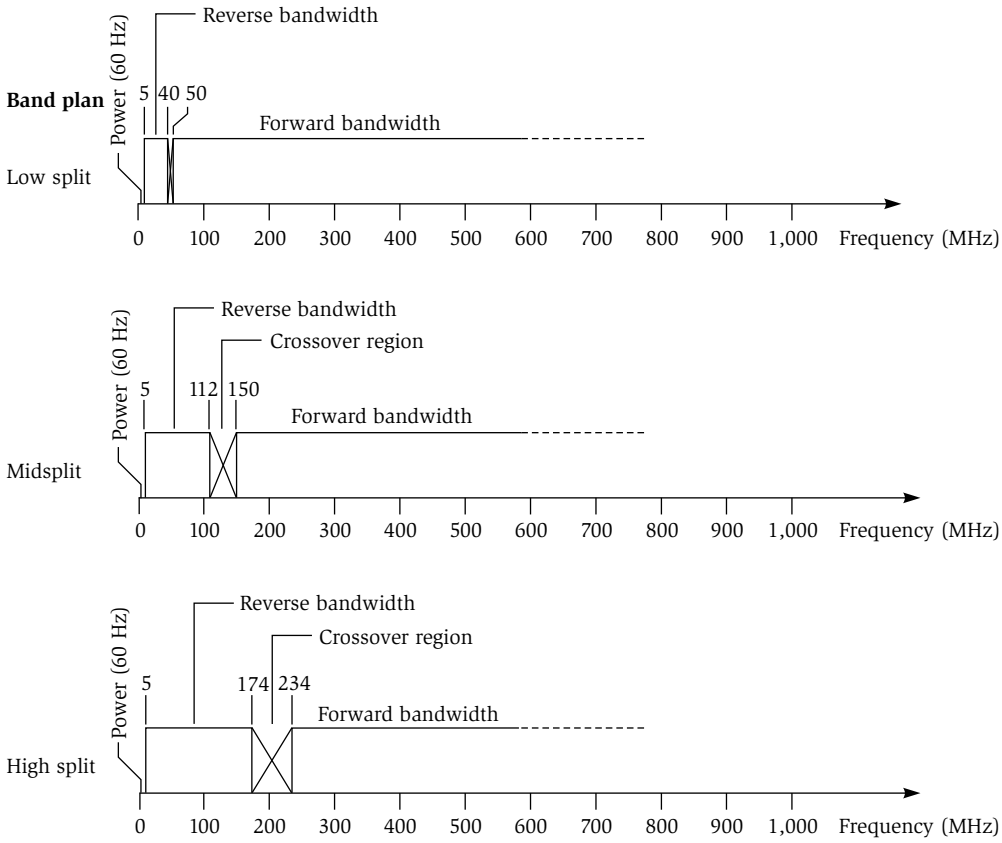


Figure 17.1 Common two-way spectrum plans.

The independence of upstream and downstream bandwidth in optical links has led to designs in which the upstream signals from each of several coaxial distribution legs emanating from a fiber node are frequency translated to nonoverlapping bands and combined before feeding the upstream optical transmitter. Known as *block conversion*, this is discussed in Chapter 16 and illustrated in Figure 16.3. The major benefit is increased average upstream RF bandwidth per customer, because signals can use the same upstream frequencies on different legs simultaneously without mutual interference.

It is common in residential system planning to divide the available downstream bandwidth among common signals, those directed to large groups of customers and those directed to individual customers. For example, it is efficient to distribute the most popular television channels systemwide, whereas local

educational or governmental access channels may vary across a large system that covers several communities, and, finally, individual cable modem data packets, on-demand movies, or digitized voice for telephony will need to be directed to specific customers. Network architectures offering various levels of segmentation will be discussed in Chapter 18. Headend requirements to create node-specific and customer-specific signals are discussed in Chapter 9.

Figure 17.2 shows the elements in a simple HFC architecture that contribute to the determination of effective bandwidth per home passed for any given service. The effective downstream bandwidth per customer for any given service in that configuration is

$$B_c(\text{fwd}) = \frac{B_s(\text{fwd})}{P_s RHN} \tag{17.1}$$

where

$B_c(\text{fwd})$ = the downstream bandwidth per subscriber to a specific service
$B_s(\text{fwd})$ = the total downstream bandwidth assigned to that service
P_s = the penetration of that service among homes passed
R = the number of nodes served from one downstream optical transmitter

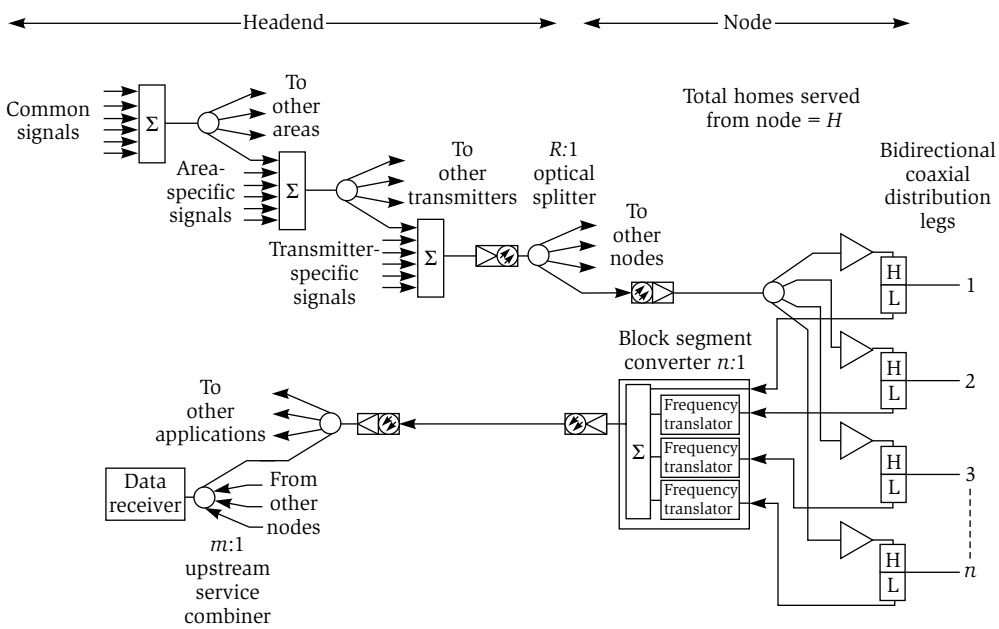


Figure 17.2 Factors controlling bandwidth per home.

H = the number of homes passed by the coaxial distribution lines extending from each node

N = the number of node transmitters fed from one service transmitter

In the reverse direction, the effective bandwidth per customer is given by

$$B_c(\text{rev}) = \frac{nB_s(\text{rev})}{mP_sH} \tag{17.2}$$

where

$B_c(\text{rev})$ = the upstream bandwidth per subscriber to a specific service

$B_s(\text{rev})$ = the total upstream bandwidth assigned to that service

P_s = the penetration of that service among homes passed

H = the number of homes passed by the coaxial distribution lines extending from each node

m = the number of nodes whose signals are combined into each data receiver input

n = the number of independent (and equally sized) coaxial distribution lines emanating from each node whose signals are effectively isolated using block converters at nodes

Note that the upstream equation is valid only if the data receiver contains an independently functioning input module for each upstream frequency block simultaneously in use.

For example, a network consisting of multiple fiber nodes, each fed by an independent optical transmitter and each providing signals to a coaxial network passing 400 homes, can provide different programming to customers in those nodes than to those fed from other, similar nodes. If the downstream bandwidth allocated to wired telephone service is 6 MHz and the service penetration is 30% of homes passed, then the bandwidth per customer would be 50 kHz. If the upstream bandwidth were 2 MHz, block conversion were used to isolate the signals from each of four coaxial 100-HP legs, and no node combining were used at the headend, then the return bandwidth per customer would be 66.7 kHz. Whether this is adequate depends on whether that bandwidth is all usable, how efficiently it is used, and simultaneous usage rates among telephone subscribers. It can be increased, however, only by choosing a smaller node size, by further subdividing the node, or by increasing the bandwidth assigned to telephony service. Such considerations have been a major factor in choosing node sizes where extensive subscriber-specific services have been contemplated.

Bandwidth Efficiency

The next important parameter is how much information can be carried in the available spectrum. This comprises two parts: information coding and modulation efficiency.

Nonscrambled analog television signals are almost always modulated in the same format as over-air broadcast signals, so direct connection to subscribers' television receivers and/or connection to those receivers after processing through simple frequency conversion set-top boxes is possible. In the United States and Canada, in fact, most cable systems are required to carry the programming making up their lowest level of service in this format. The signal formats and modulation are discussed in Chapters 2 and 9, respectively. The required bandwidth for each analog television signal is 6 MHz for NTSC and PAL-M formats and 7 or 8 MHz for PAL-B/G formats.

When video signals are converted to digital form, the resultant data rate is a measure of the information required to adequately describe the pictures and sound for the specific application. For entertainment-quality television, the Moving Picture Experts Group (MPEG) has defined a number of levels of data compression, which are described in Chapter 3. These cover resolution levels comparable to analog television as well as higher resolutions and different screen aspect ratios for high-definition television (HDTV). Since the information needed to describe a video program depends on the rate of change in the picture from frame to frame, the data rate can vary widely. On average, however, it appears that approximately 3 Mb/s is required for NTSC-quality pictures using the MPEG-2 encoding and transmission formats, whereas 19 Mb/s is required for U.S. HDTV broadcasting. Other standards, such as MPEG-4, are optimized for video transmitted through IP networks. The ratio of the digitized-but-not-compressed video stream data rate to that at the output of the compressor is the *coding gain*, which can easily be 30 or more. For nonentertainment applications, such as video conferencing and surveillance cameras, standards are available that use much lower data rates and higher coding gains.

Nonvideo data signals include packetized data for cable modems, system monitoring signals, digitized telephony of several varieties, utility control and monitoring applications, digitized audio services, interactive games and video-related services, and so on. Some of these require permanent dedicated virtual channels, while others are "bursty," with peak rates far exceeding average rates. Average data rate requirements per customer for these may vary from tens of bits per second to millions of bits per second.

Once the data stream is defined, it must be prepared for transmission. Baseband optical digital modulation is the process whereby a single datastream (which may include the time-multiplexed content from more than one source) is used to turn the output of an optical transmitter on and off. This technique is used in such formats as SONET and FDDI. While such formats have the

advantage that very high data rates can be supported (in excess of 10 Gb/s) and long transmission lines are possible, all signals must be both digital and able to time-share a common datastream.

Where digital signals are to be transmitted over a shared wideband analog network, they must first be modulated onto RF carriers. While simple on–off modulation of the RF carrier could be used, that would result in a usable bandwidth efficiency of only about 0.6 b/s/Hz. Since cable networks must provide a very high-quality signal path for analog television, more bandwidth-efficient modulation schemes are possible.

Over-the-air digital television broadcasting uses an 8-VSB format, while cable generally uses 64 or 256 QAM modulation for both data packets and digitized video (Chapter 3 covers modulation of digital video signals; Chapters 4, 5, and 6 cover nonvideo signals.) Because each transmitted symbol represents more than one bit of information, high-order modulation schemes have greater bandwidth efficiency. For instance, 64 QAM has a theoretical bandwidth efficiency of 6 b/s/Hz and a practical efficiency of about 5 b/s/Hz. Note that, after modulation onto an RF carrier, we no longer have a digital signal, but rather a digitally modulated analog signal that will share the analog transmission network with other analog signals.

The more "aggressive" modulation methods offer greater bandwidth efficiency but are less tolerant of channel impairments. The DOCSIS version 1.1 high-speed modem standard provides for up to 256 QAM for downstream transport and up to 16 QAM for upstream modulation, depending on the service tolerance for errors and channel conditions. Some upstream technologies are able to adjust their modulation order to suit channel conditions. Several proposed advanced standards, such as code division multiple access (CDMA) and orthogonal frequency division multiplexing (OFDM) would allow a more efficient use of impaired upstream bandwidth. DOCSIS version 2.0 allows use of a version of CDMA known as synchronous CDMA (S-CDMA), as discussed in Chapter 5.

In summary, the instantaneous information capacity of any part of the broadband network is equal to the RF bandwidth multiplied by the bandwidth efficiency of the signals occupying it. For example, if a network with a downstream bandwidth extending from 54 to 860 MHz were entirely loaded with 256 QAM signals (net information rate of 38 Mb/s/6-MHz channel), it could carry over 5 Gb/s of information.

The information capacity per customer is equal to the information capacity of each separate network segment divided by the number of subscribers served from that segment. In the special case of digital video, digital music, and certain voice signals, the effective instantaneous channel capacity is further related to the coding efficiency used to convert the analog signal to a bitstream. System operators will not realize all the theoretically available information capacity per customer to the extent that they use part of the bandwidth for systemwide or regional common signals.

Time Sharing

The third leg of network information capacity is how efficiently the available information capacity is shared among users and applications. As mentioned previously, traditional cable systems delivered all video channels simultaneously to all homes. In a system with many channels and small nodes, however, it is likely that some channels are not being watched in some nodes some of the time (in some small-node systems, in fact, there may be more channels than customers per node). If less popular channels are delivered to nodes only when desired by a subscriber, it is possible to offer more program choice without increasing the total available information capacity. A network able to do this incorporates a form of time sharing known as *switched virtual circuit*, whereby a user is granted full-time use of a video stream for the duration of his or her use, after which the capacity is freed up for other users. The term *switched broadcast* describes the special case where the system delivers, to one or more viewers, channels that operate on a regular schedule and are not under the control of viewers. The term *video on demand* (VOD) is generally limited to systems in which the content is streamed from a server under the interactive control of a specific viewer.

Providing a customer temporary use of a communications path is the same principal applied in the design of central office telephone switches, where use of a "path" through a switch is granted exclusively to someone making a call but then is freed up for the next caller. Depending on the statistical usage pattern, switched-channel time sharing can be a very inefficient use of available information capacity. Most telephone callers, for instance, spend roughly the same amount of time listening as talking and, even when talking, do not talk in a continuous stream but are silent for a surprising percentage of the time (though we all know exceptions!). Thus, they use only a fraction of the system resources dedicated to them during the call.

Sometimes time sharing is on a much shorter scale. For instance, a T1 telephony circuit at 1.544 Mb/s is made up of 24 digitized voice streams, each of which requires 64 kb/s (plus a total of 8 kb/s of "overhead"). Thus, 24 telephone calls are carried over one datastream, with the time slots of about 5 μs allotted to each call on a rotating basis. This is known as *time division multiplexing* (TDM); hierarchies are available up to multiple rates of gigabits per second. TDM methods that grant a fixed percentage of the total data rate to each user are known as *constant bit rate* (CBR). In the T1 example, each voice channel is a 64-kb/s CBR virtual circuit.

Where user needs are variable, more efficient sharing of a data channel results when the capacity is divided among datastreams as needed, a method known as *statistical multiplexing*. For instance, where many telephone calls share a common circuit, it is possible to greatly increase the total call capacity simply by not transmitting any bits when a caller is silent and using that capacity for another conversation. Another example of statistical sharing is the Ethernet

local area network, where computers send data packets to other computers on the network only as needed. A final, and very important, example is digital video, where the bit rate required for each video stream varies greatly, depending on the degree of change from frame to frame of a video source. It is possible to squeeze more video programs into a given datastream if we take advantage of the statistics of combining sources with varying data rates. In the case of MPEG-2 compressed, standard-resolution video, it may be possible to combine as many as 12–15 programs into a single 30-Mb/s datastream, resulting in a significantly increased program capacity as compared with analog video. In general, datastreams whose instantaneous rates vary depending on information content are known as *variable-bit-rate* (VBR) services.

Where many nonsynchronized users must share a channel, a time-sharing scheme known as *time division multiple access* (TDMA) may be used. The upstream communications from cable modems is an example of such a situation.

A key advantage of statistically sharing a digital datastream is that each application can have access to peak data rates that are much greater than needed on average. This is particularly important in the case of highly bursty data services such as cable modems used for Internet access.

Time sharing can be used at various levels in the network. For instance, entire RF channels may be shared among analog video services, contribution streams in a multichannel digital video multiplex, or individual packets in an IP network.

Summary of Information Capacity

The concept of effective information capacity per customer needs to be expanded to include statistical sharing of available capacity on two levels: simultaneous service usage rates and the instantaneous data rate required to support simultaneous users. Predictions of *simultaneous service usage rates* are done to estimate how many subscribers to a given service will use it simultaneously. This is generally known as *traffic engineering* and is well known and documented with regard to voice telephony (see Chapter 6). Usage of switched video channels and packetized data channels is less well studied. *Instantaneous circuit data rate* is a measure of the average and peak data rates needed to support simultaneously active users of VBR services.

Thus, the total information capacity of a network can be seen to be a complex that includes:

- The total RF bandwidth in each direction of the broadband distribution system
- The logical segmentation of the distribution network into subsections that can support independent signals (space division multiplexing)
- The division of the available bandwidth in each subsection among various shared and user-specific services (frequency division multiplexing)

- The division of the optical spectrum in each fiber among multiple lightwave signals, each on its own wavelength (wavelength division multiplexing)
- The coding efficiency with which signals are prepared for transmission
- The bandwidth efficiency with which signals are modulated onto RF carriers
- The statistical usage of services among subscribers
- The time-shared use of available information channels among active users (multiplexing efficiency)

Figure 17.3 shows the elements that determine effective information capacity. Not shown in this simplified diagram are such elements as forward error correction and network layer overhead, both of which detract from from net capacity.

17.2.2 Network Reliability and Availability

Service availability, in the most general sense, is a measure of the likelihood that a given service will be available for use by customers. While realiability is

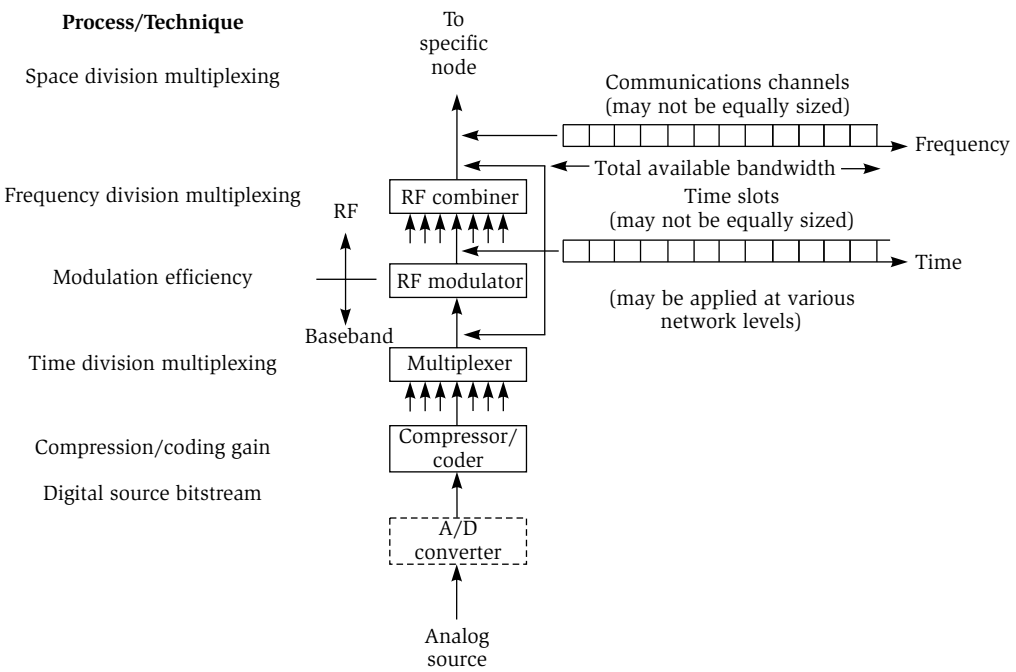


Figure 17.3 Elements that determine effective network information capacity.

related to the rate that failures occur. Some factors that affect that availability and reliability are discussed here.

Equipment Reliability

The first requirement to build a reliable network, to state the obvious, is that its component parts must be reliable. Though the discussion of internal equipment reliability is beyond the scope of this book, in general the overall reliability (usually expressed in terms of mean time between failures, MTBF, in hours) will be affected by such factors as operating temperatures, the complexity of the unit, the environmental protection, the mechanical strength of interconnections, protection from powering transients, and the long-term aging characteristics of materials and components. Typically, manufacturers provide MTBF data for their network components based on theoretical analyses and sometimes also based on test data taken under controlled environmental conditions.

Experience has shown that classical failure analyses, based on the internal electrical parts and how they are interconnected, typically yield reliability predictions that have a limited relationship to the performance of field-installed equipment. On the other hand, accurate field failure data is difficult to gather, given that field crews are often more focused on restoring service than performing postevent failure analysis.

Network Failure Rate

The failure rate of the network is determined by the number of high-level devices (such as amplifiers), their individual failure rates, and how they are interconnected, as well as certain outside factors. Chapter 20 is devoted to the calculation of network failures rates and availability. In general, however, networks that have more components that can independently cause a signal transmission failure will be less reliable. Thus, a 40-amplifier cascade will be less reliable than a 4-amplifier cascade. Not only electronic components but the power to those components contribute to failures and must be considered in the analysis. Service-related reliability requirements are one factor leading to the choice of serving-area size.

Network reliability can be improved by architectures in which the failure of some network elements does not lead to a communications outage. A simple example is a the provision of two diversely routed fiber cables between the headend and a node, with separate optical transmitters and receivers on each link and the node configured to automatically select the best signal. The probability of failing to get signals from the headend to that node then becomes the product of the probabilities of failure of each of the redundant links taken independently. For example, if each of the links has a probability of failure in some time period of 1%, then the probability of failure of the redundant network is only 0.0001 in the same time period. The gain from redundancy, however, is

reduced by the failure rate of the added equipment that must decide which of two redundant paths to use and then switch between them. Redundant structures are part of many high-reliability architectures.

In most existing cable television networks, service outages are more likely to be due to commercial electric power interruption than actual equipment failures. Thus an important factor in distribution system design is the degree and type of backup powering provided and the placement of field power supplies at points in the utility's power grid that are less likely to fail.

Network Availability—Failure Related

Given a network failure rate, whether due to component failure or powering failure, the next important parameter is the percentage of time that the network (or a specific service) is available for use. This is treated in detail in Chapter 20, but the general equation for availability is

$$\text{availability } (A) = \frac{1}{1 + \frac{\text{MTTR}}{\text{MTBF}}} \tag{17.3}$$

where MTBF is the mean time between failures and MTTR is the mean time to restore service when failures occur. MTBF is governed by the reliability of network elements and how they are interconnected, as discussed earlier; however, MTTR is both an operational practices and network design issue. Such issues as field response time and field crew training are operational issues. System monitoring capabilities, however, are part of the network design and control how soon an outage is known to repair crews. Finally, more subtle design issues, such as the number of fibers in shared cables, may affect the time required to repair a cut.

Service Availability—Signal Quality Related

Even when the physical network is operating properly, individual services may experience outages due to temporarily degraded transmission conditions. For example, an ingressing signal may interfere with an upstream data transmission. If the terminal equipment operates on a fixed frequency, the service will be lost until the signal stops transmitting or the point of ingress is located and repaired. If, on the other hand, the headend equipment is equipped to sense such interference and the upstream transmitters are frequency agile, the service can be restored by merely shifting the frequency of operation. While such considerations are not affected by distribution system physical architecture (except, of course, that smaller node serving areas are less likely to be affected by the interference in the first place), they are affected by the selection of terminal equipment and the allocation of available operating spectrum — in other words, the architecture of the specific service.

Several means of quantifying data circuit availability include signal-quality issues. Bit error rate (BER) is a measure of the percentage of individual bits of information that are lost, whereas packet error rate, errored seconds, severely errored seconds, and errored minutes all measure lost information on related but different scales. These are discussed in more detail in Chapters 5 and 6.

Service Availability—Traffic Loading Related

Where transmission paths are statistically shared among users, there is always the possibility that demand will sometimes exceed available capacity. Thus, even if the network is working properly and the transmission channels are unimpaired, the service may sometimes be unavailable to a specific user. This happens in the U.S. telephone network every Mother's Day, for instance, when, at some hours of the day, more people attempt to call their families than there are available long-distance circuits.

As with signal-quality issues, capacity-limited blockages are not affected by the reliability or architecture of the transmission network, but rather by the architecture of the individual service. In essence, they are designed as an economic choice, since providing sufficient capacity for the largest possible traffic load would not be economically justifiable if that load level occurs only rarely.

The design of an appropriate load capability for a given service is known as *traffic engineering*. This is a mature art in the case of voice telephony networks and a less well-developed art in the case of such services as cable modems and video on demand.

In summary, overall service unavailability is approximately the sum of unavailabilities due to network failures, signal degradation, and capacity-related blocking (the approximation is due to the possibility of simultaneously occurring unavailabilities in two categories).

Single Points of Failure

A final topic related to network reliability and its consequences is the management of single points of failure. Failures of components at various points in the network will affect different portions of the service and differing numbers of customers. For instance, the failure of a terminating tap will, at most, affect only the customers directly connected to it, whereas failure of the optical receiver at the node may affect hundreds of customers. Failure of a single headend channel processor may affect one channel throughout the network, whereas failure of the satellite antenna low-noise converter may take out well over 100 digital video channels.

One measure of the quality of an architecture is how immune the system is to single points of failure that affect many customers and how quickly such failures can be detected and repaired. Note that this is not necessarily the same as providing high reliability and availability. For instance, a single headend outage that takes out all services for several minutes will not affect overall yearly

availability greatly but will likely overload the customer service telephone lines with complaints, something that will not happen with a random distribution of small outages. For this reason, if no other, systems are designed to eliminate as many large single points of failure as possible. That is one reason for providing redundant structures closer to the headend, where many customers are affected, but not deeper in the network.

17.2.3 Signal-Processing Equipment Distribution

In a simple network designed to distribute a common spectrum of unidirectional signals to all terminals, the signal processing equipment can all be concentrated at the headend. This has the advantage of easy service, and it becomes simpler to design for reliability and secure powering. Where large numbers of subscriber-specific services must be delivered, however, it is sometimes more efficient to generate some of those signals, or at least to do a portion of the signal processing, closer to the users. The distribution of signal-processing equipment is an important part of architectural design. Some of the options are discussed next.

Single Point Versus Distributed Signal Processing

At some point in an HFC network, downstream signals are modulated onto RF carriers and frequency multiplexed for delivery to homes. That point may not be entirely at the headend, however. In some architectures, only commonly distributed signals are generated at the headend, and hub-, node-, and/or customer-specific signals are generated and received at lower levels in the architecture. It is always possible to create a network in which node-specific and subscriber-specific signals are generated at the headend. Moving some processing to hubs (or lower) is an economic trade-off that may also affect future network scalability. Some examples will serve to illustrate the trade-offs.

Broadcast digital video services. When the conversion from baseband digital to analog FDM signals occurs at lower levels in the architecture, the digital signals can be transported using baseband binary digital signaling to the point where they are modulated onto RF carriers. This allows longer optical spans that are free of the types of mutual interference that limit the performance of linear distribution techniques. The trade-off is that it is more expensive to convert from digital to RF modulated signals at each hub than simply to pass signals through without format conversion.

Interactive digital video services. When signals are spooled from local storage, as is the case with VOD, moving some of the servers to hubs reduces the instantaneous information capacity required between headend and hubs, because each increment of content need be moved only once to the server. The trade-off is that the total load cannot be balanced across

servers, so more may be required. Also, any solution that puts complex equipment at hubs requires either continuous staffing at those points or the acceptance of longer service restoration times when outages occur.

Private communication services. In the case of services whose content is merely transmitted by the network to and from specific customers (for instance, Internet access data and telephone calls) it may be efficient to put the subscriber interface equipment (the CMTS for data or the HDT for telephone) at hubs, because the network side of both types of equipment can be multiplexed and share a high-speed digital link with other hubs and the headend.

As the demand for data capacity grows, it may be more economical to place the first level of data multiplexing even closer to the user. Several manufacturers offer weatherproof high-speed data multiplexers that are designed for mounting at the curb level with separate baseband drops to each user and optical network-side connections from one or more higher-level multiplexers. Depending on the design, this data layer may be a complete parallel overlay on an existing HFC network (sharing only optical cable sheaths) or may share some communications links and/or powering. An example of such an architecture is discussed in Chapter 18.

Point-of-Entry Versus User-Terminal Signal Processing

At the customer end of the link, it is possible to distribute broadband signals directly to terminals at the point of use (and to generate any required upstream signals from those same points). Doing so, however, has several distinct disadvantages. These are most apparent in two-way systems:

- There may be unnecessary duplication of functions among receivers and transmitters within a single building.
- Electrical transients or over-air signals can leak into the system anywhere within the building and pollute the upstream spectrum of the entire node.
- Although equipment from multiple manufacturers and transmission standards can be accommodated, these must all compatibly use the available spectrum.
- Network powering of terminal equipment, though possible, is much more difficult because many buildings are wired by other than network operators and because many drop passive components are not designed to handle multiplexed power.
- Due to the many possible in-building distribution system topologies, the available downstream signal strength and required upstream transmitter levels will vary widely.

The alternative is to terminate and generate some or all network signals at a single *point-of-entry* box located where the drop cable enters the house. This is

sometimes also known as a *network interface device* (NID) or *network interface unit* (NIU). Typically, but not always, analog video signals are passed through the NID to customer terminals, whereas digital signals of all types are terminated at the NID and reprocessed for distribution through the house. Similarly, all upstream signals may share a single upstream transmitter at the NID.

The advantages to use of a NID include:

- Isolation of the network from the effects of in-building wiring, both losses and signal ingress
- Simplified network powering, particularly for wired telephony services
- Cost savings in multiservice installations through sharing of common equipment

Balanced against these advantages are the generally higher costs of the first installed service due to the flexibility required in the NID to support incremental features or services. Also, the use of a common NID implies, in most cases, that all two-way services must be interoperable. Often in the cable industry, this means a common and sole source for equipment.

In summary, network architecture includes not only the physical network topology but the distribution of signal-processing equipment for various services. Network information capacity can vary by service and includes the RF bandwidth, how efficiently that bandwidth is used and shared, and the effects of equipment distribution on the information required to support the service.

On an operational level, architecture has an effect on the reliability and availability of services, on the size and frequency of outages, and on how difficult it is to recover from those outages.

17.3 Requirements by Service Type

A key factor in designing a network architecture is a clear understanding of the services that will be offered, because each has unique requirements. The signal quality requirements for various analog and digital services are covered in detail in Chapters 2–6 and will only be summarized here. Additionally, specific services require different levels of reliability and availability, switching, per-customer data capacity, and links between network customers and networks that are external to the cable system. Although many service categories may not be covered in the following discussion, the general process of assessing requirements should be similar.

17.3.1 Broadcast Video Services

The U.S. cable television industry was founded on the retransmission of analog, NTSC, over-air broadcast service to improve reception for viewers as compared with what they could get with private antennas. In the late 1970s, those channels

were supplemented by satellite-delivered networks. In the context of this chapter, *broadcast services* will refer to all programming delivered to all subscribers in common. Today, most of the bandwidth of typical residential networks is still dominated by analog programming delivered in nonscrambled, broadcast analog format to subscribers. The requirements for such networks are that they have sufficient downstream bandwidth for all the channels, coupled with a quality and availability that is acceptable to subscribers (and that conforms with regulatory requirements, as outlined in Chapter 15).

With the practical implementation of digital video compression, the theoretical program capacity of such networks has been increased by 6 to 10 or more times for standard-resolution programming and by 2 times for high-definition programming, compared with over-air broadcast formats. Digitally modulated video signals, whether in broadcast (VSB modulation) or cable (QAM modulation) format, have a greater tolerance to noise and intermodulation distortion than analog video signals (of the order of 30 dB C/N for 64-QAM modulation), with the result that they can be run at lower power levels in the network and, therefore, provide less loading of the active devices.

As opposed to the lifeline telephony situation, where network reliability is related to emergency services access and is, therefore, a governmental and public concern, most video services consist of nonessential entertainment, and therefore network reliability is primarily a marketing issue. In other words, the programming must be delivered with sufficient reliability that customers do not prefer alternate service providers. It can be argued that the transmission of Emergency Alert System (EAS) warnings through cable networks creates an overriding public interest in their reliability, but no federal standards have been issued (though local governmental authorities sometimes negotiate customer service standards with franchised cable operators that include outage measurements).

Cablelabs' Outage Reduction Task Force reported on industry research to determine customer tolerance of various outage rates and durations in their landmark 1992 publication *Outage Reduction*.[1] Based on its findings, that group proposed the following performance target:

> No customer should experience more than two outages within a three-month period. While less accurate, because the cumulative effect of outages over time is lost, a secondary standard could be set at no more than 0.6 outages per customer per month.[2]

This guideline is based on research that showed that, when outage frequency exceeded the target rate, it became a major factor in customers' perception of service quality, whereas below that rate, the overall satisfaction rating was largely independent of outage frequency.

Another important result of the research showed that, though there was a difference in perceived quality for outages of various lengths, the dissatisfaction

level for 10-minute outages was nearly as high as for all-day outages, and the dissatisfaction at outages of 1-minute duration was about one-half the 10-minute rating. The task force, however, did not put forward a recommended standard for service restoration, or network availability percentage, though they state that "A[n availability] figure in a good month is 99.99% or greater. In a poor month, the figure might be as low as 99.95%."[3] If we use an average service restoration time of 4 hours for all outages, then 0.6 outages per month corresponds to a network availability of 99.7%, a frequently quoted cable industry availability goal.

In the same document, the task force has proposed the following definition for an *outage*:

> An *outage* is defined as any event in which two or more customers experience loss of reception on one or more channels arising from a common cause, regardless of the cause. *Loss* is defined as in interruption rather then degradation of signal. Loss of a single channel at the headend or hub site is included.[4]

This is an interesting definition from a network design standpoint. For one thing, it includes only outages "experienced" by subscribers. Thus, outages that are detected and repaired before they are noticed by subscribers do not count. The relationship between actual failure rate and subscriber-experienced failure rate is discussed in Chapter 20.

Second, by excluding single-subscriber outages, it ignores all outages occurring because of faults in drop wiring or terminal equipment, whether furnished by the network operator or the subscriber. Given that many, if not most, subscriber outages arise in either the drop or terminal equipment, this is an important exception.

By excluding single-subscriber outages, the definition also eliminates from consideration any single-channel outages that affect only single customers. Though this is not important for broadcast services, it is important in the case of interactive services such as video on demand, discussed later, where each video stream is utilized by only a single subscriber.

Clearly, the outage definition is not consistent with the research on subscriber tolerances. Subscribers are likely to be just as dissatisfied with an outage occurring because an individual set-top converter failed to descramble as one occurring because a headend modulator failed. It would be prudent for those designing networks to include all elements involved in service delivery in setting acceptable outage rates. Given the relative independence of outage length on dissatisfaction rating, it would also be prudent to set separate goals for outage rate and network availability (or service restoration time).

A final parameter in determining reliability parameters for delivery of broadcast video services is outage size. When customers experience an outage, a certain percentage of them will attempt to contact the network operator to report the event. If the outage affects many customers, this can cause the telephone call volume to swamp the customer service department and its incoming telephone trunks. Since many franchises (as well as voluntary industry guidelines) contain

requirements regarding telephone responsiveness, it is incumbent upon operators to design networks that have as few common elements as possible and to tailor repair efforts to respond most quickly to those outages affecting the largest number of customers.

Although individual network operators must set their own service standards, a reasonable goal for broadcast video service levels is an outage rate not exceeding 0.6 per month, as defined earlier, but including single as well as multiple customer outages. Independent of the outage rate, network availability should be at least 0.997, corresponding to an average restoration time of 4 hours combined with 0.6 outages per month. Service restoration times should be managed so that they are inversely proportional to the number of customers affected.

17.3.2 Narrowcast Video Services

Narrowcast services are defined, for purposes of this discussion, as those that are not available to all subscribers but whose generation is not under direct control of viewers. Included in this category are the services known as pay, pay-per-view (PPV), and near video on demand (NVOD). In each case, the services are delivered, in accordance with a fixed transmission schedule, to a subset of the subscribers receiving the broadcast channels.

The three categories blend into each other. Pay services historically refers to services such as HBO that are subscribed to on the basis of monthly fees. PPV refers to a fixed schedule of movies and events that are sold on a per-event basis, with different orderable events transmitted sequentially on one or more channels. NVOD is a logical extension of PPV whereby movies are typically displayed with sufficient frequency that the average wait time for a particular movie is short enough to discourage subscribers from seeking other means to access the programming.

The parameters specifically related to narrowcast services include the means used to control access, ordering mechanisms, and the efficiency of network information capacity usage. Access can be controlled by either signal denial or signal scrambling. Scrambling and signal denial methods are discussed in Chapter 21 and will not be further covered here.

Ordering mechanisms include:

- Manual advance ordering by telephone (person-to-person to a CSR)
- Manual advance ordering by telephone using the telephone key-pad dialer to enter the information into an audio response unit (ARU)
- Manual or automatic advance ordering using automatic number identification (ANI) technology
- "Store-and-forward" systems whereby advance access is granted to programs with post-event accounting of purchased events
- Real-time ordering of programs

Store-and-forward systems typically use a downstream data carrier to poll individual descramblers on purchases. Depending on the model, the individual descramblers may use a telephone modem and the public switched telephone network to upload information to the computer or, in the case of a two-way cable system, may transmit the information upstream through use of an RF transmitter. Real-time ordering systems involving two-way transmission capacity have an advantage over use of the telephone because they allow the operator to avoid paying for multiple telephone trunks and modems and also to avoid problems when customers' phone lines are busy.

Whether store-and-forward or advance authorization is used, the upstream data requirements are modest — just sufficient information to identify the subscriber and programs involved in the transaction. Transmission latency requirements for store-and-forward systems are very lenient — hours or even days being acceptable in some cases. In real-time ordering systems, customers may tolerate a few seconds to a couple of minutes between requesting access to a program and its enablement.

It has been shown through numerous marketing studies that the ease with which programming can be accessed is a critical parameter in controlling average buy rates. It has also been shown that, despite incentives for advance ordering, subscribers tend to order programming immediately before needing it, so order-processing rate is also a critical parameter.

As systems offer an increasing number of optional video services, the effective information of the capacity of the network devoted to those services must be increased. There are basically three ways to do that: increasing downstream bandwidth, providing greater information density (for example, digital video compression) and/or implementing switched broadcast technology. The first two are obvious, but the third has not been widely exploited as an alternative.

Common Versus Switched Information Capacity

Given a need to deliver a defined range of services to a group of customers, operators have the option of providing sufficient network-wide information capacity that all signals can be delivered throughout the network or providing switching and multiplexing so that a reduced instantaneous information capacity, combined with a customized signal load for each network segment, accomplishes the same end. In multilevel architectures, this switching can take place at the headend or at some lower level or levels in the network.

In the special case of switched broadcast services using digital video, several variables need to be considered: how many RF channels will be used, how many subscribers will share a set of switched broadcast signals, what the cost per stream is to switch the input signals and what the cost is to create and modulate each output multiplex. The optimal engineering solution will depend on the specific problem. Switched broadcast architecture is discussed in more detail in Chapter 18, Section 18.2.5. Figure 17.4 illustrates the principles and hardware involved.

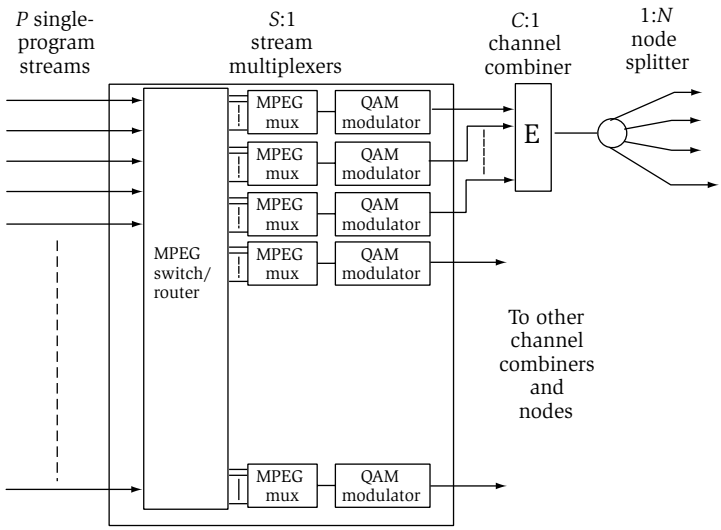


Figure 17.4 Switched broadcast parameters.

Neither the U.S. cable industry nor federal regulators have suggested reliability standards for optional video services, as opposed to broadcast services. The research done by CableLabs, discussed earlier, did not distinguish among service levels or consider the relative expectations of subscribers to premium services. Since customers pay more per channel for such services, though, it is logical to predict that they would expect a higher reliability for such services.

17.3.3 On-Demand and Interactive Video Services

On-demand video services are defined as those delivered to specific subscribers, generally in response to real-time requests. Such programming might include more than entertainment categories. Generally this implies a degree of real-time interaction between the subscriber and the program generation and control system through a two-way cable system. As an example, a subscriber might use a multilevel menu system to select from among audio, video, and print material or might search a database of articles in a research project. Alternatively, the subscriber might select a movie and start time from an available library, and then be granted VCR-like control over its playback (known as video on demand, VOD).

From an architectural requirements standpoint, on-demand services will differ primarily in the acceptable latency times and in the nature of the information source. In the case of video games, for instance, acceptable latency times may be of the order of a few milliseconds, whereas database searches may tolerate several seconds, depending on the application.

Whatever the nature of the information being transmitted, the downstream requirement is to provide exclusive use of a communications channel for each simultaneous user. As with narrowcast services, required bandwidth is affected by service penetration rate, simultaneous usage rates, and node size. The upstream requirement is for a channel with appropriate bandwidth and latency. Generally, the upstream bandwidth per user, even for the extreme case of interactive games, is relatively modest, though latency requirements may be strict.

With respect to reliability of service delivery, it can be assumed (because of the date) that the CableLabs research on subscriber tolerance of outages was based largely on broadcast and monthly services. As with narrowcast services, it is reasonable to assume that subscribers who are sole users of channels and who may be paying far higher per-channel costs for that exclusivity will be more demanding of service reliability. On the other hand, narrowcast services are viewed for a shorter time than, say, broadcast services, so the statistical probability of failure will be lower, a balancing factor. To the extent that these narrowcast, on-demand services share the same physical cables as the broadcast video channels but also use relatively complex server devices, their inclusion in a network operator's offerings may place a more stringent reliability requirement on the broadband delivery system.

17.3.4 Packet Data Services—Internet Access

For purposes of this discussion, *packet data services* are defined as nonvideo digital transport in which data is packaged in short bursts and sent through a time-shared channel on a demand basis. The most common noncable example of such a network is an Ethernet, in which variable-length packets are sent between any two terminals on the network. Included within this broad definition are both TCP/IP and ATM.

Packet data services may be offered occasionally through a one-way cable network (typically with telephone modems used for the upstream packets) but more commonly through a full two-way network. The network quality requirements are those that will result in an acceptable service quality. This means that the net loss of packets (after any provided forward error correction) must be low enough that the net throughput does not materially suffer. Typical, well-maintained HFC networks will provide uncorrected bit error rates in the range of 10^{-9} to 10^{-10} in the downstream direction with 64-QAM modulation and 10^{-5} to 10^{-7} in the upstream direction with QPSK modulation.

In the downstream direction, a 6-MHz RF channel containing a 64-QAM modulated signal will provide a data rate of about 30 Mb/s; 256 QAM will carry 38 Mb/s. The total number of such channels required will be a function of service penetration, simultaneous usage among subscribers, average data rate per simultaneous user, acceptable blocking rates, and the number of homes that share common data signals.

In the upstream direction, the requirements are slightly more complex. In addition to the preceding factors, the upstream capacity includes the efficiency with which users are controlled so as to share each RF channel. A common scheme requires modems to request use of the channel and then to be allotted time slots on a specific upstream RF channel, as discussed in Chapter 5. In order for the system to accommodate subscribers who might be located at different distances (and therefore different transmission times) from the headend and at different tap ports, upstream transmitters must be level-agile and able to synchronize their transmissions in time.

Although average data rates will change as usage patterns and applications develop, a typical cable modem traffic analysis might go as follows: Each simultaneously–active user's session is assumed to include:

	6	downstream messages/minute
each consisting of	100	packets
with an average length of	1600	bytes/packet
multiplied by	8	bits/byte
divided by	60	seconds/minute
equals	128	kb/s *average* data rate per active session

If we assume that the dominant use of the modems is for file downloading, then the upstream traffic will consist primarily of 64-byte acknowledgments of the 1,600-byte downstream packets, a 1:25 ratio. Thus, the upstream data rate to support the given downstream rate is 128 kb/s ÷ 25 = 5.12 kb/s.

Note: To the extent that continuous streaming and home-server-based applications proliferate, these suggested average rates may be low by an order of magnitude or more and upstream and downstream rates may become more symmetrical.

It is common to market data services based on the peak gross downstream user rate, with the upstream limit being set to a fraction of that — typically 20–25%. What the user experiences, however, may have little relation to what is marketed. Among other factors:

- The user-experience data rate will always be lower than the peak rate, due to DOCSIS IP overhead.
- The peak rate may exceed the advertised rate unless the network operator programmatically sets the maximum rate.
- The peak rate may be less than that advertised when usage of the network is heaviest and there is contention for resources, both on the RF and network sides of the CMTS.
- The measured rate to any Internet site will be limited by the capacity available on every intervening link in the path.
- Finally, the rate will be limited by the remote server and contention among simultaneous users there.

The relationship among node size, number of nodes sharing a datastream, usable downstream data rate, service penetration (measured against homes passed), and simultaneous usage rate is given by the following:

$$\frac{\text{usable data rate}}{\text{average data rate per active session}} = \text{HP/node} \cdot \text{nodes} \cdot \text{penetration} \cdot \text{simultaneous usage}$$

where the usable data rate is the net available user data rate in the data stream; average data rate per active session is as calculated earlier; HP/node is the number of homes served from each node; nodes is the number of nodes that share a downstream data stream; penetration is the percentage of homes in each node that subscribe to cable modem service; and simultaneous usage is the number of data subscribers using the service simultaneously.

For example, if we assume that the net usable data rate from a 64-QAM, 6-MHz-bandwidth signal is 20 Mb/s and that the average downstream data rate per active user is 128 kb/s, then each channel will support 156 simultaneous users. If two 1,000-home nodes share a common downstream signal and we allow for up to 35% simultaneous usage among data subscribers, then the system will support a 22% penetration among homes passed.

In the upstream direction, total usable data rates, data rates per session, and circuit sharing will differ; otherwise the calculation is similar.

Since the upstream spectrum is subject to ingressing signals (and also malfunctioning modems that transmit continuously), many modem systems are also upstream frequency agile, so the system can "sidestep" a degraded frequency. Enabling that capability, however, requires that additional bandwidth be allocated.

Network reliability/availability standards have not been set for packetized data services. To the extent that such services compete with those delivered by dial-up telephone modems, the local public switched telephone distribution network arguably will meet the 0.9999 availability recommendation for local telephone service. Balancing the possibly-higher cable network failure rate, however, is the ability of cable modem equipment to "ping" (test) each modem on an ongoing basis and thus to quickly determine if there is a communications or equipment problem. Even a relatively low penetration of such devices will ensure that virtually the entire cable network can be continuously monitored for integrity. By contrast, the telephone company knows of an individual line problem only when the subscriber detects it and manually calls it in.

Although telephone company physical networks are generally reliable, the accessability of the local Internet service provider (ISP) may be far less certain. A 1997 survey by Inverse Network Technology, for instance, found that the call-in failure rate for all measured ISPs was 22.4%, whereas the largest provider in the country had a failure rate of over 60%![5] By providing permanent virtual connections between users' PCs and the Internet, network operators using cable

modems will be able to offer a considerable availability enhancement over dial-up alternatives.

For another competitive comparison, AT&T's Digital Data System (DDS) which offers duplex data channels to users at speeds ranging from 2.4 kb/s to 1.544 Mb/s, guarantees a BER of at least 10^{-7} and an availability of 0.9996, with the local circuits on each end of the network (T1 carrier plus baseband loop) assumed to have an availability of 0.9999.[6]

17.3.5 Voice-Grade Telephony Services

Voice-grade telephony comes in several forms, including conventional wired telephony, PCS telephony, cellular telephony, and IP telephony, also known as *voice over IP* (VoIP). Wired telephony requires a physical connection between the user's terminal equipment and the distribution system, whereas PCS and cellular systems use bidirectional radio transmission. VoIP can be supported over any IP network with sufficient transmission fidelity and latency properties. Although there are distinct legal and frequency usage differences between cellular and PCS telephony, from a network perspective there is almost a continuum of terms of cell sizes, cell interconnection requirements, and available services. Generally, cellular base stations and telephones both use higher power than PCS and correspondingly greater spacing between base stations. Also, PCS is a purely digital service, whereas cellular can be either analog or digital in the air link.

Wired telephony service can certainly be provided over an HFC network, as can the interconnection between PCS transceiver antenna locations, though the network requirements are considerably different. Despite promising early engineering work, PCS station interconnect has never been widely deployed. However, the interested reader is encouraged to consult the Endnotes for further information.[7–9] VoIP is an emerging technology that shares DOCSIS-based transmission equipment with computer data. The details of the DOCSIS standards are discussed in Chapter 5.

Switched Circuit, Constant-Bit-Rate Telephony

By definition, wired telephony requires that the cable systems support two-way communication to and from every home. Since the greatest percentage of upstream ingress problems occur in the drop system and internal building wiring, the communications equipment used must be designed to work in relatively "dirty" channels. One architectural choice that eliminates many in-home problems is the use of an external NID on each home that contains the data receivers and transmitters for telephony, with standard twisted-pair copper used to transport baseband analog telephony within the building. By using a NID, the network is protected from in-home ingress (see Section 17.2.3 for a general

discussion of the advantages and disadvantages of the use of NIDs). NIDs may be telephone-only or, with VoIP, may terminate both voice and data services.

The requirement to have the service survive localized power outages also favors the use of a NID, which can be directly powered from the coaxial plant, which in turn can be equipped with standby power. In other cases, operators may choose to equip each NID with batteries and a local charger. Either choice involves compromises. Network powering, unless the NID has very low power consumption, can significantly increase the total power consumed by the network and for that reason make the design of reliable standby power more difficult. The use of a home-powered charger and batteries creates the need to maintain that equipment and to dispose of worn-out batteries on a regular basis. Also, the power for the charger within the home is likely to be less reliable than distribution plant power. Finally, though it is common for telcos to plan for long-term outages by equipping standby plant power supplies with generators or very high battery capacity, that is not economical on a home-by-home basis, with the result that individuals will be protected only for short power outages.

Telephony signals are sent to and from individual subscribers. Thus, the RF bandwidth used will be related directly to node size, number of nodes sharing common telephone signals, modulation efficiency, service penetration, and simultaneous usage among subscribers. As an example, assume that a system is built with 500-home nodes, with pairs of nodes sharing telephone signals. If the network operator is a local exchange carrier (LEC) and expects to serve virtually all residences with at least a primary telephone and a 50% chance of a second phone, then it must plan on a capacity of 1,500 lines per service group. Assuming the company's traffic engineering studies predict 20% maximum simultaneous usage (known as *line concentration* in the telephone industry), then the network must be able to handle 300 simultaneous telephone calls per node. At 64 kb/s per voice stream, the network must be able to transmit 19.2 Mb/s per node. Finally, if QPSK modulation is used on the RF carriers (with a bandwidth efficiency of 1.5 b/s/Hz), the theoretical required bandwidth will be 12.8 MHz in each direction. Commercially available equipment is not that bandwidth efficient, however, and is typically limited to 24 simultaneous calls per 2 MHz, so a total of 24 MHz would be required, equivalent to four standard 6-MHz downstream television channels and a significant percentage of the usable upstream bandwidth of most standard low-split systems.

As the primary communications link into homes, the local telephone service is expected to provide very high reliability and availability and to work when the power is out. All of these are greater than the demands required to deliver nonessential entertainment video services. On the other hand, some feel that there is a viable business in providing lower-cost secondary lines whose requirements are less strict.

BellCore, the research consortium of the former Bell operating companies, has established reliability objectives for the telephone industry. Those relating to the subscriber loop are contained in TR-NWT-000418.[10] The basic criterion is

that "the two-way service availability objective of a narrowband transmission channel should be a minimum of 99.99% (0.9999 probability) for the subscriber loop." This corresponds to an average yearly unavailability of 53 minutes. *Narrowband transmission channel* is further defined to mean services, including voice, that are transmitted at speeds lower than DS-1 (1.544 Mb/s).

It is important to understand what the BellCore criterion includes and what it does not. In terms of applicability, everything in the central office on the subscriber side of the switch (or remote terminal) is included. For HFC systems, the equivalent headend components would include the RF modulators, for instance, but not the shared HDT. BellCore's definition includes the distribution network and, if used, the NID. The subscriber-provided in-building wiring or telephone instrument is not included.

Equipment failures and the time to repair them are included in the calculation. Not included, however, are service outages due to both primary and backup powering. Since the majority of outages in HFC networks are power related, actual subscriber unavailability may far exceed 53 minutes per year and yet meet the BellCore criterion. By comparison, the CableLabs guideline quoted earlier includes power outages. Also, BellCore's definition defines the restoration time as beginning when the operator becomes aware of the outage, which may be long after it actually starts in the case of an outage affecting a single customer. Finally, the unavailability experienced by an individual customer is further affected by his or her usage pattern and by failures of his or her own terminal equipment. Despite its limitations, the BellCore criterion is accepted for measuring availability of residential lifeline telephone service.

The relationship of subscriber-perceived outage rate and unavailability to network reliability and availability is complex and involves usage patterns as well as the depth of network monitoring. This subject is discussed further in Chapter 20. It is interesting that, though the added complexity of HFC networks using NIDs may *increase* the actual hardware failure rate, the addition of automatic subscriber line monitoring may allow a *decrease* in customer-perceived outages relative to lower-failure-rate, but unmonitored, networks.

VoIP

IP telephony began as a means of bypassing the long-distance carriers. Computers at each end of an IP telephone link were equipped with microphones, speakers, and audio coders/decoders (CODECS) and could use voice communications, provided the two-way data rate was sufficient to support the digitized voice.

Although the initial results were crude in comparison to the normal public switched telephone network, hardware improved to the point that major carriers have taken an interest in it. One development has been the emergence of facilities in numerous cities that serve as gateways between the internet and local exchange carriers (LECs). This allows calls between standard phones and

IP phones. Both the legal and technical frameworks for this service are still under development.

When IP telephony becomes commonplace, it will change cable modem service usage patterns, including session timing and length and the degree of upstream/downstream data rate symmetry. The DOCSIS 1.1 and PacketCable standards were developed, among other reasons, to encourage the development of VoIP by ensuring interoperability of various manufacturers' equipment.

17.3.6 Utility Monitoring and Control Services

A final example of services that can use the broadband infrastructure are various low-average-data-rate monitoring and control functions. Examples are electric utility customer load control and meter reading; burglar, fire, and medical alarm monitoring; and traffic light monitoring and control.[11] These services can be characterized generally as relatively low-speed data circuits with varying latency and transmission reliability requirements. Most require highly accurate end-to-end data integrity.

The two applications with the greatest apparent hope of commercial success are electric meter reading and demand-side management (DSM). Since electric power companies must provide capacity based on peak, rather than average, demand, DSM offers them a means of trimming peak electric loads by selectively turning off high-powered appliances, such as hot water heaters and air conditioners, during the hottest hours of the day. By turning off rotating groups of such appliances for a short time, they can materially reduce peak load while minimizing inconvenience to users.

Although the data rates are not high, the overall reliability by which the DSM process functions must be exceptional, because the result of failure could be a widespread power outage. On the other hand, only the common equipment must meet this high standard, because the failure of a single node or of the equipment in a single home will not have a material effect on overall load management. Nevertheless, it has been suggested that network availabilities in excess of 0.99997 might be required by some utilities.[12] By contrast, the requirements for meter reading are less stringent, since timing is not critical and data can be retransmitted if necessary to ensure accuracy.

A similar situation exists for medical, burglar, and fire alarm monitoring. The average data rates are quite small, and transmission delays of a few seconds are tolerable. But network availability is crucial, as is transmission accuracy, since lives could potentially be lost if alarms are not promptly and accurately reported.

17.4 Scalability

Engineering has, somewhat simplistically, been defined as finding the lowest-cost solution to a technical problem, given a complete definition of that

problem, including performance, operational parameters, and reliability. Unfortunately, in broadband network engineering, the requirements are usually very ill defined.

The network's owners (be they stockholders, partners, or investors) want a network that provides an initial set of services to customers, with adequate quality and reliability and at the lowest cost. However, they also expect the network to be able to provide undefined future services when and where required and to be able to provide services to whatever penetrations of homes marketing is able to sell — all this, if possible, without abandoning any of the original capital.

Thus, scalability has become an important element of network architecture. It can be measured by the following parameters:

- The ability to increase the bandwidth per subscriber for any given service without increasing the cost of providing other services (*service-specific scalability*)
- The ability to increase the bandwidth per subscriber in a given geographic area without increasing the network or service provision cost elsewhere (*area-specific scalability*)
- The ability to increase the bandwidth per subscriber for any given service in a given geographic area without increasing the cost of providing other services or the same services to other areas (*service- and area-specific scalability*).

Chapter 18 presents in detail an example of a specific architecture that will illustrate the trade-offs required to achieve scalability in each of these parameters.

17.5 Summary

Networks can be described by a number of parameters, including:

- Noise, distortion, and other signal-degradation measures
- Instantaneous RF bandwidth
- Various means, such as segmentation, that increase the effective bandwidth per user
- Reliability and availability
- Service-specific and geographic-specific scalability

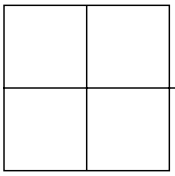
Services that may be delivered over those networks place varying demands on the distribution system:

- The quality of the RF channel occupied by its signals
- The occupied bandwidth, both forward and reverse
- The availability of the network to carry signals
- Network powering for unique service-related processing equipment

The network architecture must support adequate performance and effective bandwidth for all contemplated services. Chapters 18 and 19 will discuss various architectural examples and how they meet the needs of certain services; Chapter 20 will deal with reliability and availability modeling.

Endnotes

1. Cable Television Laboratories, *Outage Reduction*, September, 1992, Cable-Labs, Louisville, CO.
2. Ibid., p. I-18.
3. Ibid., p. V-10.
4. Ibid., p. I-8.
5. As reported in: David Plotnikoff, Dialing Up Answers from the FAQ File. *San Jose Mercury News*, June 15, 1997, p. 6E.
6. Bernhard E. Keiser, *Broadband Coding, Modulation, and Transmission Engineering*. Washington, DC: CEEPress Books, 1994, pp. 96–97.
7. George Hart, Cost-Effective Cable Television Transport for PCN. *1992 NCTA Technical Papers*. National Cable Television Association, Washington, DC, May 1992, pp. 148–156.
8. Douglas E. Holulin and Robert W. Hammond, Integrated Wireless PCS — Hybrid Fiber Coax Network Architecture. *1996 NCTA Technical Papers*, National Cable Television Association, Washington, DC, April 1996, pp. 10–18.
9. George Hart, op. cit.
10. Bell Communications Research, *Reliability Assurance for Fiber-Optic System Reliability and Service Availability*. TR-NWT-000418, Issue 2, December 1992.
11. Rainer Kochan, Dipl. Ing., Value-Added Services and Functions for MATV and CATV Systems, *Symposium Record, CATV Sessions, 16th International TV Symposium*, Montreux, Switzerland, June 1989, pp. 524–535.
12. Dean Ericson and Dan Carter, Demand-Side Management Considerations in Advanced Network Deployment. *1994 NCTA Technical Papers*, National Cable Television Association, Washington, DC, May 1994, pp. 8–14.



Chapter 18

Architectural Elements and Examples

18.1 Introduction

Network architecture is the logical and physical interconnection of all the elements between a signal's generation and its termination. Chapter 17 dealt with ways of measuring architecture-related parameters and the needs of various types of services. This chapter will first delineate each of the elements from which an architecture is constructed and then present some examples of high-level HFC distribution networks that have been proposed or constructed. Fiber-deep architectures will be covered in Chapter 19; the relationship between network architecture and service reliability and availability will be dealt with in Chapter 20.

18.2 Architectural Elements

Although network architecture is often thought of in terms of the physical distribution topology, everything from the internal headend (and preheadend) processing through network termination equipment affects how services are delivered. It is important, therefore, to understand the implications of available options at all levels of the network. In this section we discuss only those structures that are part of the cable television network; we do not include information flow outside that system.

18.2.1 Terminal Equipment

The network terminal equipment provides the interface between the user's end application devices and the shared broadband distribution system. Depending on the service, it may perform format-conversion, testing, or security functions as well as protecting the network from disruption. One example is set-top converters that perform tuning and selective descrambling functions for video services and may also provide menuing or ordering capabilities. In the case of digital video channels, the box also converts the signal to standard analog video

format. In two-way systems, the box communicates user's service requests to the headend. Another example is a telephony point-of-entry network interface device (NID), which converts between digitally modulated RF signals and analog baseband telephone signals for routing within a home. Choices in terminal equipment can have a significant effect on overall cable system design and performance.

In-Home Equipment

Historically, the full cable spectrum (possibly minus some channels blocked by traps) has been distributed throughout subscribers' homes. In many cases, subscriber-owned television receivers have been connected directly to cable outlets; in others, converters or converter/descramblers have been placed between the drop and the subscriber's receiver.

This scenario is consistent with U.S. regulations that allow customers to wire their own residences and that require cable operators to deliver all channels in at least the lowest television programming tier in unscrambled, analog broadcast format.

One advantage to delivering the full cable spectrum throughout homes is that any service can be accessed from any outlet by connecting the proper equipment. Second, no home equipment investment is required in advance of service delivery. Finally, the cost and responsibility of equipment powering are borne by subscribers.

Balancing those advantages are the relatively high and variable loss of drop wiring in homes, which makes it more difficult to ensure that terminal equipment receives the proper signal levels. Additionally, home wiring has been a major egress point for cable-carried signals and ingress point for broadband noise, signals, and electrical transients in two-way systems. Although upstream transmitters may be located anywhere within the home drop system, the full bidirection bandwidth must be preserved, and upstream plant is therefore exposed, on average, to more frequent interference. Often cable operators reduce this risk by installing filters to block upstream ingress on portions of the drop system that will not be used by two-way devices. Finally, and depending on the services being offered over the cable system, there may be a duplication of some functionality among different pieces of terminal equipment.

Point-of-Entry (POE) Equipment

Terminating the broadband linear network at a common point exterior to the residence provides a means to isolate two-way cable networks from signals generated or picked up within the home. Any upstream signals to be transmitted into the distribution system can be generated by a common transmitter at the POE device. Similarly, multiple downstream services may share some of the common POE circuitry. For example — the POE bidirectional transceiver may

generate and receive standard baseband signals for transmission over twisted-pair cables to telephone instruments and drive an Ethernet bus or wireless transceiver serving multiple computers while transmitting and receiving RF signals from the headend.

Where several services are provided and can be coordinated, the use of shared POE equipment can result in both cost savings and improved operation. The addition of automated status monitoring may be justified by improved monitoring capabilities. On the other hand, the cost of equipment for the first service delivered to a home is often more expensive than a single-purpose in-home version. Finally, lacking interindustry standards, use of a multiservice POE device restricts the ability of operators to buy equipment from multiple sources.

There is a long-standing debate regarding the relative advantages of network versus home powering of POE devices. Home powering significantly reduces the power consumption of the distribution network, but it is less reliable. Also, it is much simpler to maintain large shared standby powering systems than individual in-home units. The goal to provide 99.99% availability for at least basic telephone service certainly tends to favor network powering. The proper decision is less obvious for nonessential services.

Shared Terminal Equipment

In some cases, terminal equipment is integrated with multiport taps, so some functions (for instance, power supplies and data transceivers) can be shared among multiple homes. The most common examples for video services are addressable disconnect functions, interdiction, and "whole-house" descrambling systems (these are discussed in Chapter 21).

Two obvious advantages of shared terminal equipment are easier maintenance access and, potentially, improved security. Deployment of such a system, however, requires either that the network provider completely equip the plant with such equipment or that the plant be modified each time a new subscriber is installed. In other words, the operator is forced to choose between a high initial capital cost and an even higher evolutionary upgrade cost accompanied by service interruptions.

As with POE equipment, shared terminal equipment has the further disadvantage of limiting future flexibility — it is hard to upgrade program scrambling security when the descramblers are part of the distribution system. These problems have limited the widespread adoption of shared interdiction technology by the cable television industry. The greatest commercial success with shared equipment has been in multiple dwelling units (MDUs), where the rate of connects and disconnects exceeds the average, signal theft is high, and/or physical access is difficult.

Shared terminal equipment is central to fiber-to-the-curb (FTTC) architectures. In an FTTC architecture, the fiber distribution network extends to curbside shared terminal equipment, whose functionality is shared among a few homes or businesses (typically 6 to 25). Individual homes are fed from the terminal

equipment through fibers, twisted-pair lines, or coaxial drop cable. Fiber-deep architectures are treated in depth in Chapter 19.

Hybrid Arrangements

Most commonly, where either POE or shared terminal equipment is used, it is coordinated with some in-home equipment. For example, several manufacturers have announced POE network interfaces for telephony and data that pass the bandwidth containing video signals to set-top descramblers. The upstream bandwidth is divided, with data and telephony signals generated by the interface equipment and signals related to ordering of television signals originating in the set-top box. Filters block only a portion of the upstream bandwidth from the home wiring. Though such arrangements are arguably more expensive than totally integrated solutions, they preserve the network operator's flexibility to deal with different vendors and formats and work well where new services are to be installed without disrupting older service platforms. Figure 18.1 illustrates terminal equipment location options.

18.2.2 Coaxial Distribution Network

The generic design of coaxial distribution systems was covered in Chapter 11. Within the general framework of amplified, split, and tapped networks, however, there are many architectural variables.

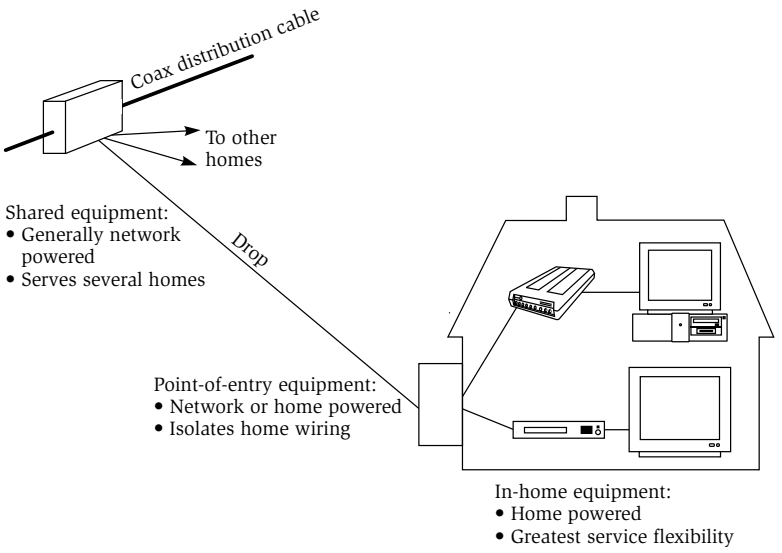


Figure 18.1 Choices for network termination equipment location.

RF Bandwidth

The most basic parameter of any coaxial network is the forward and return bandwidth. Since these are determined primarily by the amplification equipment, some operators are proposing the use of coaxial distribution networks small enough that they can be totally passive between the output of the fiber node and the user. Such networks have the advantage that the division between upstream and downstream bandwidth can be altered at will, as can the upper frequency limit. As a practical matter, however, these freedoms are restricted by the spectrum usage issues discussed in Chapters 9 and 17, and the upper frequency will be constrained by the characteristics of the passive devices used, by the increasing loss of the cables, and by the practical limits on signal levels in the network, as discussed in Chapter 10. The eventual conversion of all video transmission to digital and the cessation of VHF over-air broadcasting, of course, will enable major changes to cable system spectrum allocations.

Coaxial Serving Area Size

The next most basic decision in the design of an HFC network is the size of the area (measured in number of homes passed) to be served by one physically-separate coaxial network. Since the use of the RF bandwidth is shared among all the users in the serving area, this determines the upper limit on bandwidth per home passed.

Since reliability is inversely proportional to the number of series-connected critical components, the required network availability and failure rate for the services to be offered should also be a factor in determining serving area size. The calculation of network reliability and availability is discussed in detail in Chapter 20.

Typically, three to four separate coaxial distribution legs will be served from each fiber-optic node, offering opportunities for subsequent node size reductions. Some operators, in fact, design nodes for subsequent subdivision.

Coaxial Powering

The average reliability of commercial power is often lower than the total end-to-end required network performance. Thus, systems are required to employ a powering strategy that minimizes the effects of power interruptions. Conventional cable television "standby" power supplies typically provide 1–2 hours of operation from self-contained batteries. While this is sufficient for short interruptions, some percentage of interruptions will exceed this length. In systems where customers are affected by more than one such power supply, reliability will further decrease due to the typically noncorrelated nature of those interruptions.

Where the individual coaxial distribution legs are limited to approximately 150 homes passed (assuming typical suburban, single-family home densities), it

is practical to power the entire coaxial subnetwork from a single location, even when network-powered POE interface devices are used. This is, in fact, the strategy typically used by U.S. telephone companies who have constructed HFC networks. Single point powering of this size of network, however, generally requires using 90 volts, with maximum currents of the order of 15 amperes. Voltage limitations as well as electrical safety issues are controlled in the United States by the National Electrical Safety Code or local codes, at the option of each community.[1]

In order to achieve telephone-grade reliability, operators offering primary-line telephone service and high-availability data services have generally equipped these common power supplies with 8-hour battery capacity and/or backup generators. The generators, in turn, can be powered by self-contained fuel supplies or tied to available natural gas mains.

In order to avoid the problems and costs of installing such power supplies at each node, one telco (Southern New England Telephone) proposed powering nodes from the central office. To accomplish that, they developed a hybrid cable containing optical fibers for signals plus copper conductors to carry 440-volt, three phase power. At each node, the power was converted to acceptable voltages for multiplexing onto the coaxial distribution systems.[2]

Where the current-handling capabilities of components or the voltage drop through distribution cables might otherwise limit application of single point powering, the output from one power supply is sometimes divided among several power distribution cables whose resistance per unit length is less than the distribution cables. These power bypass cables may be routed alongside the signal cables past the first split point, where lower currents can be inserted into each of several network legs. This is illustrated in Figure 18.2. Since these power

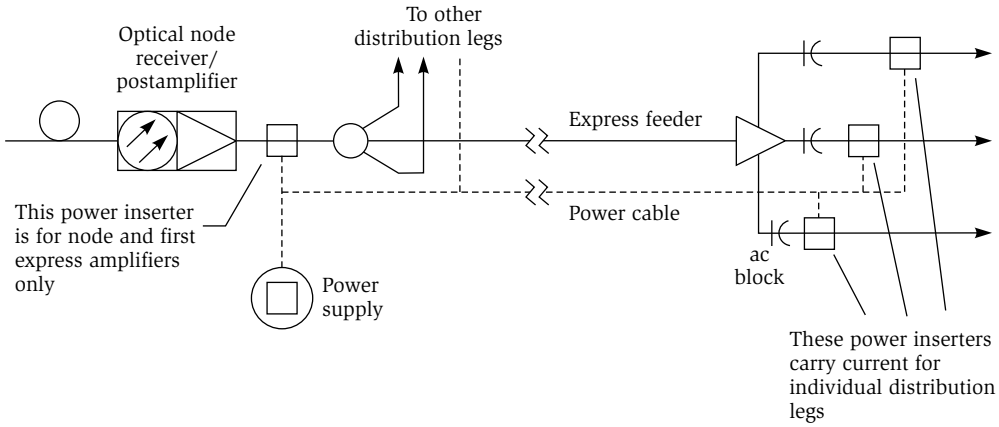


Figure 18.2 Bypass powering to increase reach of node-located power supply.

distribution cables do not carry RF, lower impedances and nonstandard construction are sometimes used to reduce power loss.[3]

When battery backup supplies are used, major improvements in outage rate and availability can occur if the supplies are equipped with monitoring systems that alert network operators of commercial power outages so that crews with portable generators can provide interim power before the batteries expire.

18.2.3 Fiber Node

Individual fiber nodes provide the interface between the linear fiber trunking system and coaxial distribution legs. At the simplest level, such a node may consist of a single optical receiver whose output is amplified to feed the downstream amplifier and, optionally, an upstream optical transmitter whose input is driven by the output of a combiner whose inputs are the upstream signals from all connected coaxial distribution legs. There are, however, many architectural variations.

Split-Band Multiple Fiber Inputs

Some nodes provide the option for multiple downstream input fibers that are fed to separate optical receivers whose outputs are then combined in a diplex filter. There are basically two advantages: improved signal quality and network flexibility. In the first case, the C/N and distortion of the fiber-optic links can be improved by simply dividing the total signal load between two transmitters. This allows a greater optical modulation index and thus an improved C/N. At the same time, fewer distortion products will be produced. The relative gains in C/N and distortion can be exchanged, within a limited range, by adjusting the optical modulation depth.

A more common reason to use split-band optical inputs is to allow those signals that are common to multiple nodes to be carried on one fiber and those that are node-specific to be carried separately. As will be seen later in the chapter, this can result in overall network flexibility and cost savings. One should be cautioned, however, that a system that uses split-band inputs cannot easily adjust those bands, because the diplex filters in every node would have to be changed.

Redundant Multiple Fiber Inputs

An alternate use for multiple downstream input fibers is redundancy of the fiber-optic trunking network. In nodes equipped with redundant inputs, each fiber is routed to its own full-bandwidth receiver. A switch at the output of the receivers selects the better (or surviving) signal in the case of failure or degradation of one path. If the fibers feeding the node are diversely routed (so that the probability of simultaneous cut is low), then significant gains in reliability can result. As a

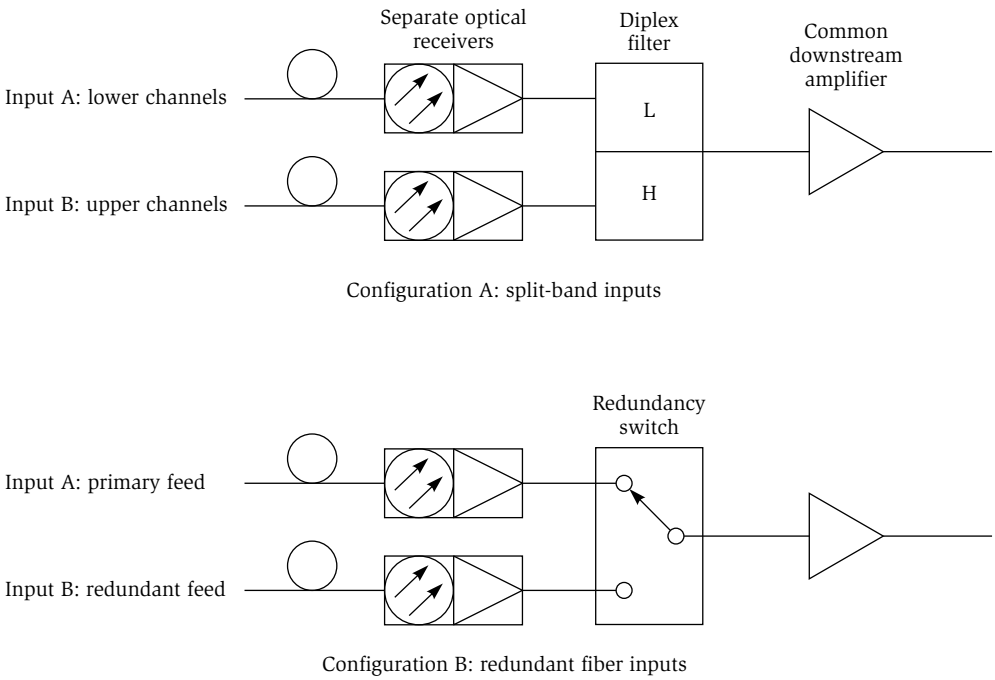


Figure 18.3 Fiber node multiple input options.

practical matter, the gain in reliability is limited by the reliability of the switch (and its control circuit), which is an added critical element in the signal path. As Figure 18.3 shows, the configurations are very similar; in fact, some manufacturers offer either as an option in the same node platform.

Redundant Signal Processing

For applications where availability is critical, nodes are available with additional redundant elements. Power supplies, downstream and upstream amplifier modules, and upstream optical transmitters are all available in dual configurations. In the last-mentioned case, dual upstream optical outputs are provided, enabling diversely routed external cables.

Handling of Multiple Coaxial Distribution Legs

Generally, optical nodes are used to feed more than one coaxial distribution leg. Although this is a design convenience that allows for shorter coaxial cascades by placing the feed point near the logical center of the node-serving area, electri-

cally separating the coaxial legs also offers opportunities for effective bandwidth expansion and improved upstream performance. One option, described in Chapter 16, is block segment conversion, whereby the upstream signals from different legs are converted to nonoverlapping frequency ranges. Thus the bandwidth per home passed is limited by the coax-serving area rather than by the larger node-serving area. Offsetting the per-subscriber bandwidth gain is the increased performance required of the upstream optical transmitter, which must handle more signals and increased RF bandwidth. In some cases, this will require upgrading from FP to DFB transmitters, at a considerable increase in cost.

An alternate way of achieving the same upstream bandwidth expansion is to use separate upstream transmitters and fibers for each distribution leg. This reduces the performance requirements for each transmitter, but requires more transmitters and fibers. A third alternative is to use separate upstream transmitters operating on different optical wavelengths and to combine their outputs using a wave division multiplexer (WDM). Whether separate fibers or WDM is less expensive will depend on the length of the fiber transmission path and the relative cost of the transmitters and multiplexers.

A final option is first to process the entire upstream spectrum from each coaxial leg by converting it to a high-speed baseband digital signal, to time multiplex these signals, and then to feed them to a high-speed baseband digital optical transmitter. Several manufacturers offer versions of this technique, with either two or four inputs multiplexed to create a single baseband datastream of about 2 Gb/s. These transmitters are also available in ITU-grid DWDM wavelengths, so they can be further multiplexed at the hub before transmission to the headend.

These four options are illustrated in Figure 18.4. Note that optical splitting or use of additional laser transmitters can, in theory, be used with any option to create dual outputs for feeding diversely-routed fibers for higher upstream reliability, though the incremental cost is least for the single-fiber configurations.

18.2.4 Fiber Interconnects

Linear fiber-optic links provide the connectivity between the point where the FDM spectrum is created and its delivery to coaxial distribution legs. Fiber links are also used at higher levels in the network to distribute selected signals of all types, often in digital form. Although fiber-optic links are used among headends and between headends and outside information sources (such as broadcasters and Internet service providers), this section will discuss only those links that form part of the network extending from headends to and from end customers.

Although this book will treat interconnect architectures as if they were all constructed with fiber optics, it should be remembered that amplitude modulated broadband microwave radio technology (AML), as discussed in Chapter 14, is still used for both individual signal and broadband distribution in those cases where it is more cost effective than construction of physical optical cables.

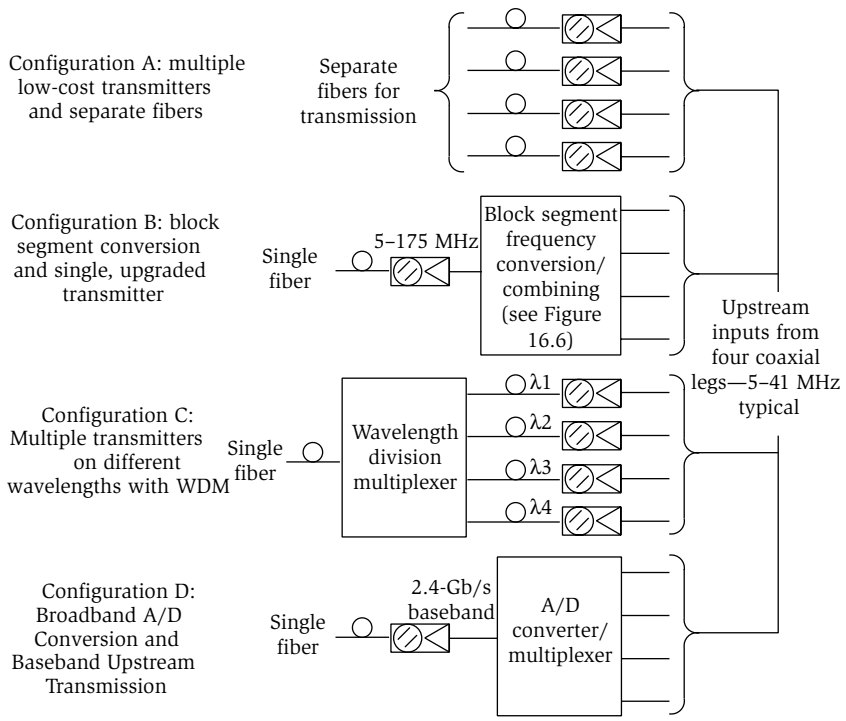

Configuration A: multiple low-cost transmitters and separate fibers
Separate fibers for transmission
Configuration B: block segment conversion and single, upgraded transmitter
Single fiber
5–175 MHz
Block segment frequency conversion/ combining (see Figure 16.6)
Upstream inputs from four coaxial legs—5–41 MHz typical
Configuration C: Multiple transmitters on different wavelengths with WDM
Single fiber
Wavelength division multiplexer
λ1
λ2
λ3
λ4
Configuration D: Broadband A/D Conversion and Baseband Upstream Transmission
Single fiber
2.4-Gb/s baseband
A/D converter/ multiplexer

Figure 18.4 Options for segmentation of upstream coaxial inputs.

Clearly, however, microwave links can replace only point-to-point fiber links and not more complex fiber topologies. Also, the available microwave spectrum may not permit as many channels as in a fiber-linked system. For instance, the U.S. 13-GHz CARS band will support only 80 analog television channels.

Star

The simplest optical architecture, and the one most commonly deployed for transmitting analog FDM signals from headends or hubs to and from nodes, is the *star*, defined as separate paths from a common point to multiple termination points. When constructed with dedicated optical fibers, driven by separate transmitters, the star architecture allows complete independence in the signals delivered to each node (as opposed to the most common type of microwave broadband star network, which splits the output of a common transmitter to feed multiple antennas and paths). In fact, many cable systems optically split fiber transmitter outputs to feed two to three nodes, but do that splitting in the

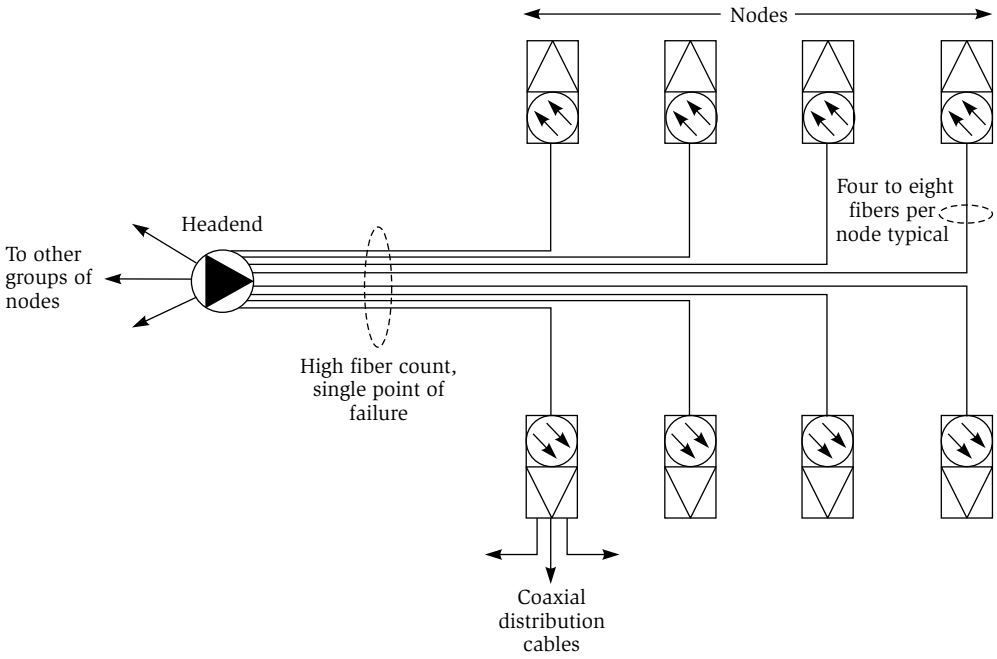


Figure 18.5 Star architecture.

headend so as to allow easy separation of nodes in the future when greater network segmentation is appropriate.

Though the failure of an individual fiber feeding a receiving point affects only the subscribers on that node, practical cable routing often results in one cable sheath holding the fibers for many nodes and thus creates a possible single point of failure that is much larger (see Figure 18.5).

Sheath Ring

The exposure to outages caused by fiber cuts or failures can be reduced significantly by sending the signals between the common point and each termination via physically diverse routes. When multiple termination points can be arranged in a ring configuration, considerable savings result from the sharing of common construction costs, as shown in Figure 18.6. This architecture is known as a *sheath ring*. Note that the diagram shows only a single line in each direction around the loop from the common point to each node, whereas in fact there may be several (at least one upstream and one downstream). Analysis of the figure will show that the cable can be cut at any point without destroying continuity from the common point to each node.

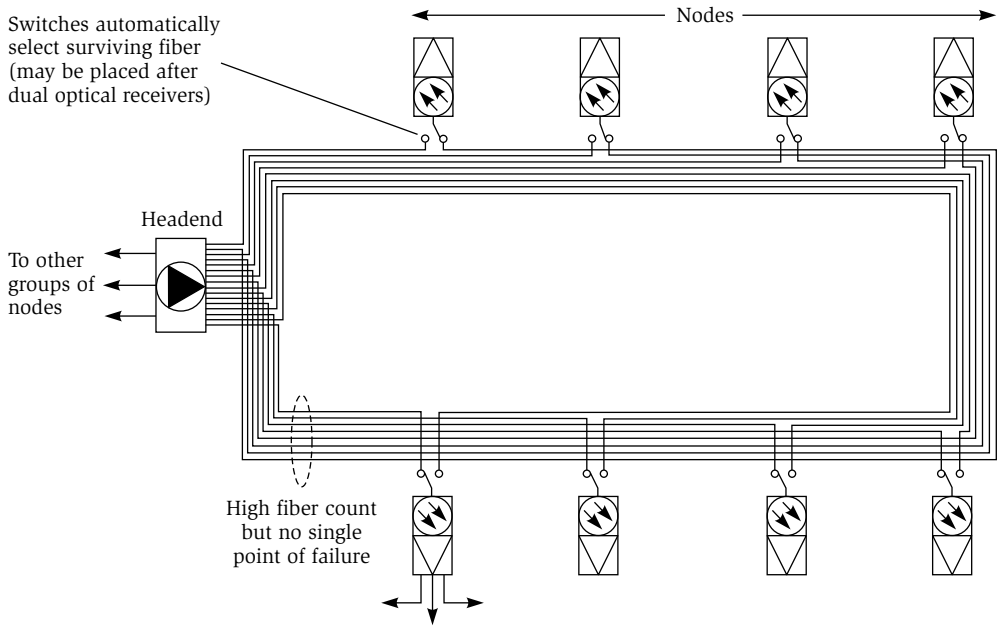


Figure 18.6 Sheath ring architecture.

Although the sheath ring architecture preserves the signal independence of the star and adds a considerable degree of protection against fiber cuts, it does so at the expense of a considerable increase in total *fiber* footage. Depending on the physical layout of the ring, however, the total *cable* footage may be only slightly greater than for a simple star. Since construction costs typically are much greater than fiber costs, the incremental cost of the installation may be only moderately higher.

Sheath rings are a common architecture among competitive access providers (CAPs) who offer highly reliable data transport to commercial customers. They are also commonly provided between headends and hubs in multiple-level architectures, to protect against potentially very large outages should a major optical trunk cable be cut.

Analog Shared Ring

Where a common set of signals must be sent to multiple nodes, an economical structure is the shared ring, illustrated in Figure 18.7. The common signal spectrum is transmitted both clockwise and counterclockwise from the common point using separate fibers (and often separate transmitters to increase the degree of

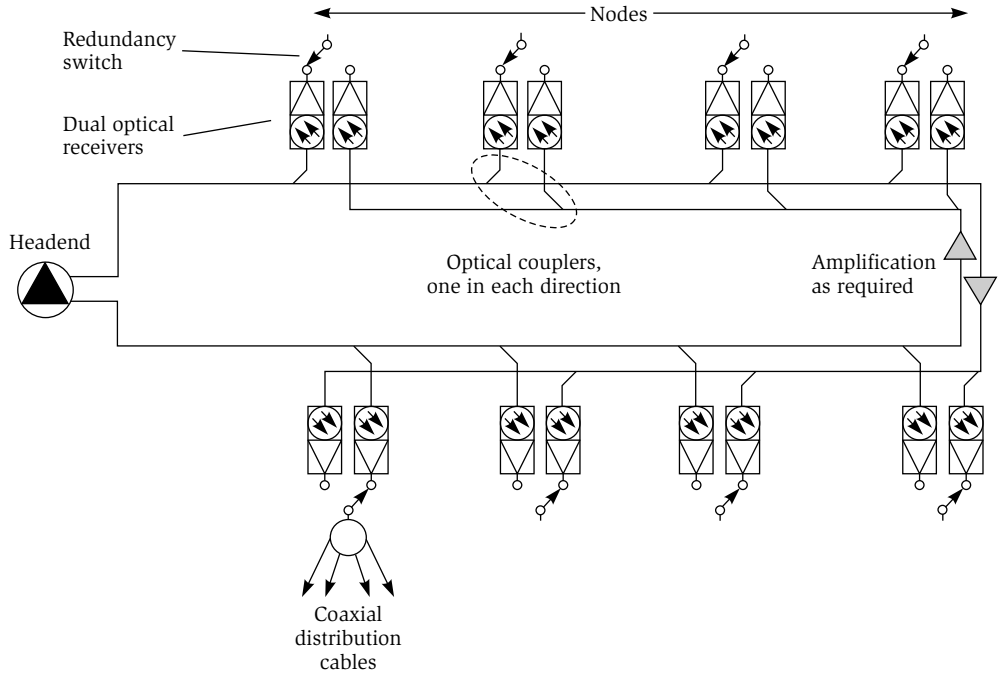


Figure 18.7 Analog shared ring architecture.

redundancy). At each node, a portion of the signal is coupled from both rings and fed to separate optical receivers. A redundancy switch, as described in the optical node options, selects the surviving path in the event of a cable cut or transmitter failure. Alternatively, the optical switch can be placed ahead of a single receiver.

Though shared rings can be constructed at 1310-nm wavelength, the maximum available optical budget of approximately 17 dB (see Chapter 12) limits the number of receiving nodes. At 1550 nm, however, fiber losses are less, and optical amplifiers are available that can considerably increase the available loss budget.

A major advantage of the shared ring is obviously a considerable savings in optical fiber usage and number of transmitters, as compared with either a simple star or a sheath ring, while preserving the protection against fiber cuts and transmitter failures. Although its application is limited to distribution of a common signal spectrum, shared analog rings are effective when combined with other structures.

Digital Repeating Ring

A ring structure that is well suited to baseband digital signal distribution is the repeating ring, an example of which is shown in Figure 18.8. As discussed in

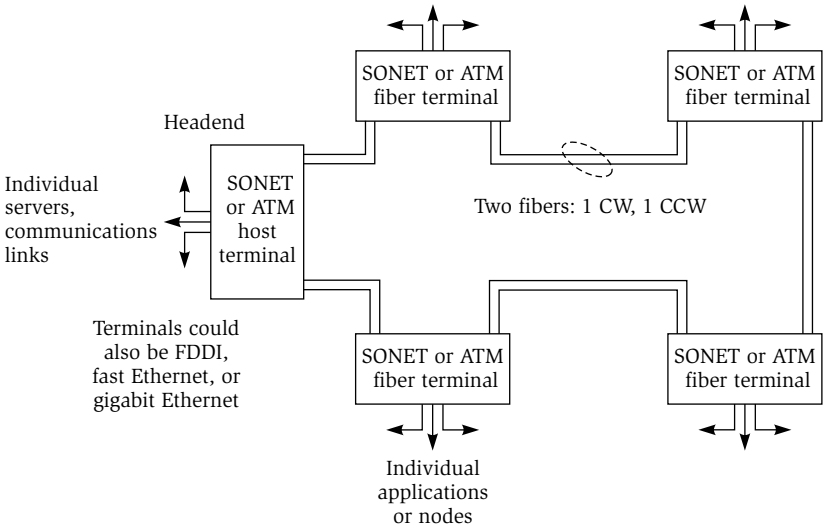

SONET or ATM fiber terminal
SONET or ATM fiber terminal
Headend
Individual servers, communications links
SONET or ATM host terminal
Two fibers: 1 CW, 1 CCW
Terminals could also be FDDI, fast Ethernet, or gigabit Ethernet
SONET or ATM fiber terminal
SONET or ATM fiber terminal
Individual applications or nodes

Figure 18.8 Baseband digital repeating ring.

Chapters 12 and 19, baseband digital optical transmission allows much greater optical budgets (of the order of 30 dB) and very high modulation rates.

In a repeating ring of the form shown, the information destined for all termination points (usually hubs) is time multiplexed into a single datastream, which is transmitted both clockwise and counterclockwise around the ring. At each hub, the signals destined for customers served from that hub are separated from the stream and signals destined for the headend are inserted. Should a cable be cut or any piece of terminal equipment fail, the remaining equipment automatically "loops back" so that it communicates with the common point via the surviving path.

As with the shared ring, an advantage to the architecture is the small number of fibers required while preserving protection against cuts and equipment failures. A limitation is that its application is restricted to the distribution of information that can be converted to the common digital transmission format. Several such formats are available, including proprietary time division multiplexed (TDM) schemes, asynchronous transfer mode (ATM), synchronous optical network (SONET), and gigabit Ethernet. When used to transport NTSC video, digital transmission requires costly signal processing at the point where the signals are converted to RF, because each must be demodulated to a baseband digital stream, converted to an analog baseband signal, remodulated to the desired RF channel, and, finally, combined into the desired FDM spectrum for delivery.

A video transmission variant that is lower in cost than SONET and designed specifically for the cable television market uses pulse code modulation (PCM). Rather than starting with a baseband digital signal, it encodes the modulated IF signal for transmission. Then, at the destination, it is necessary only to reverse the process and convert the reconstituted IF signal to the desired RF output channel.

Depending on the format, digital terminal equipment may take the form of switches, routers, or add–drop multiplexers. Where standards-based formats such as SONET are used, network capacity can be upgraded by simply using terminal equipment that supports higher data rates on the network side. Rates up to 10 Gb/s are commercially deployed.

18.2.5 Switched Video Headend Structures

In traditional entertainment-only cable systems, the headend served the function of receiving signals from various locations and processing them into a common spectrum for distribution throughout the distribution network. As discussed in Chapter 9, considerable complexity is added by the requirement to generate custom spectra for each node containing both regional differentiation and information to be transmitted to and received from individual customers.

An important example that illustrates the architectural trade-off between node size, RF bandwidth, and frequency reuse is the use of headend switching in the offering of near video on demand (NVOD) — typically a selection of movies played on a fixed schedule with frequent start times — or narrowcast (meaning low viewership) video services. The characteristics of both service categories are that the video sources operate on a fixed schedule not controlled by the viewer.

If a common spectrum of signals is distributed throughout a cable system, then the number of program offerings is obviously limited by the spectrum devoted to that service classification and the efficiency with which that spectrum is used.

The usage of video channels, however, is very nonuniform, with the result that some channels are not always being watched in all nodes. For example, if a network is built with 1,000 home nodes, subscriber penetration is 60%, and simultaneous usage (percentage of homes watching television at any one time) is 67%, then 400 viewers will be watching at the peak hour of usage in any one node. Based on current viewing patterns in North America, the major television networks command approximately 50% of the viewing audience. Including those, the total percentage of customers tuned to any of the 20 most popular channels on any cable system is likely to be over 90%. Thus, regardless of the number of channels offered, the maximum number of additional channels that will be watched at any one time in any node is less than 40 and will approach that number only if the viewers of less popular programming all want to view something different from each other. Put another way, any channel whose audience share is less than 0.3% has a less than 50% likelihood of being accessed in any given node at any given time.

These statistics suggest an alternative way to provide channels of limited interest. If viewers' channel requests could be communicated to the headend, then a video switch and separate set of modulators for each node (or multinode group of customers) could be used to provide less popular channels only as required. If this structure were put in place, it would remove any limitation on the number of program offerings except for the economic limitations on switch size. Such an analog architecture was tested successfully in a U.S. cable system, though it was not widely deployed.[4] A digital video variant is now being offered by a different manufacturer. This version has the dual advantages of more efficient use of bandwidth and the ability to utilize the two-way set-top box communications infrastructure already provided for interactive video services, such as VOD.[5] The principles are illustrated in Figure 18.9.

The switched NVOD/narrowcast service differs from a true video-on-demand (VOD) service in the assumption that all available NVOD and narrowcast channels are always present at the input of the matrix switch, whereas in the

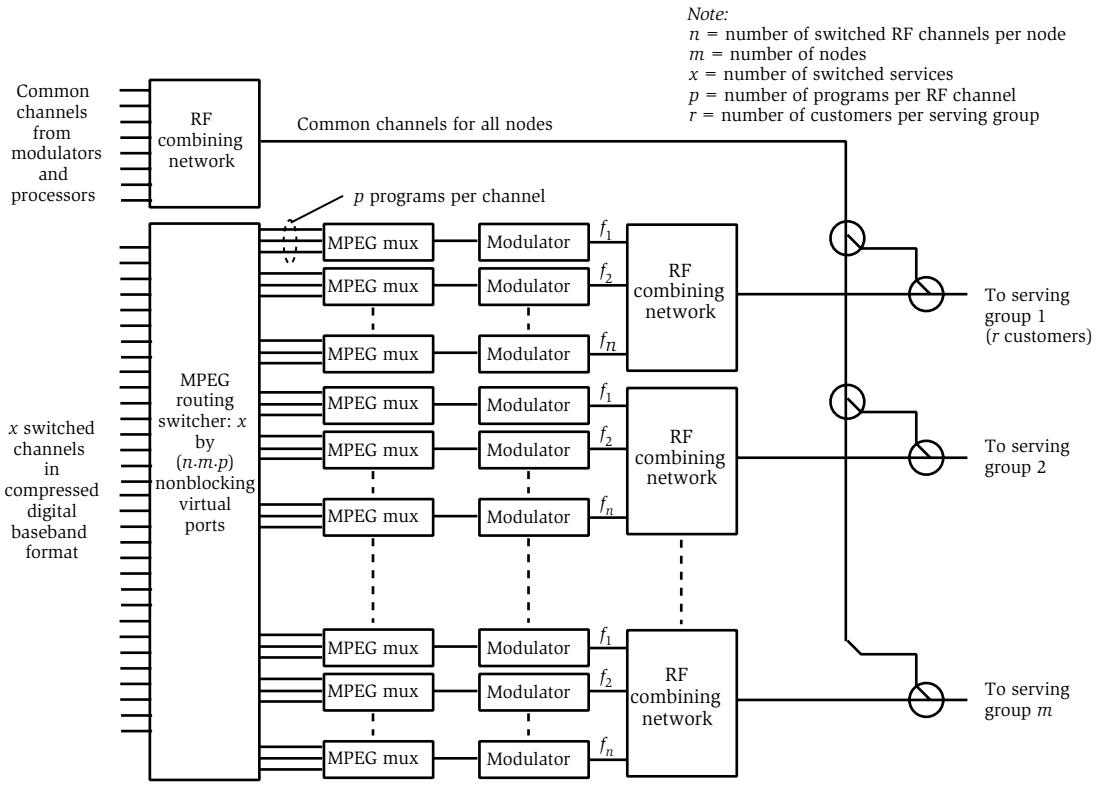

Note:
n = number of switched RF channels per node
m = number of nodes
x = number of switched services
p = number of programs per RF channel
r = number of customers per serving group
Common channels from modulators and processors
RF combining network
Common channels for all nodes
p programs per channel
MPEG mux
Modulator
RF combining network
To serving group 1 (r customers)
To serving group 2
To serving group m
x switched channels in compressed digital baseband format
MPEG routing switcher: x by (n.m.p) nonblocking virtual ports

Figure 18.9 Switched NVOD or narrowcast headend functional diagram.

VOD case, channels are generated only by the server in response to (and under control of) individual subscriber requests. In the application described, video is not *generated* in response to subscriber request but only *switched*, so it is available to subscribers fed from specific nodes. Thus, the result is *switched broadcast*, not interactive VOD.

The trade-offs in deploying switched broadcast services are complex.

- *Serving Group Size (r in Figure 18.9)*. Smaller serving groups will each need fewer multiplexers and modulators (n) because the number of switched channels required will be statistically less, though the relationship is not linear. Balancing that, the number of sets of multiplexers and modulators increases linearly with the number of serving groups (m), which is inversely proportional to serving group size.
- *Division between common and switched bandwidth*. The more channels that are offered through the switched platform (x), the higher the statistical likelihood that a customer on a given serving group will request a switched channel (thereby increasing the number of modulators per serving group), though, again, the relationship is not linear.
- *Statistical sharing of modulators among nodes*. At the expense of additional RF switching hardware and complexity, frequency agile modulators could be shared among nodes, in recognition of the fact that the number of subscribers requesting switched channels will vary randomly with time. That would require replacing the RF combining networks with switching matrices and managing both the frequencies of the modulators and their assignment to nodes.
- *Sharing of channels among subscribers*. When more than one subscriber in a given node requests a switched channel, the system can simply supply the program on two channels or, at a slight increase in complexity, can direct the second requestor to the already supplied program. This capability protects the system against overload when an occasional offering on a switched channel is unusually popular.

As an example, consider an NVOD service that offers the top 10 movies, each starting every 15 minutes. Given a 2-hour average movie length, this would require 80 program streams if every event were to be available to every subscriber. At a 10:1 net compression ratio (that is, 10 digitized, compressed movies transmitted in the spectrum generally used by a single analog channel), 8 RF analog channels would be required.

If this offering were made using a switched broadcast architecture, then we need to make some traffic assumptions. Assume that each service group includes 1,000 homes. At a 70% basic service penetration, the group will include 700 video customers. If 50% of those choose to lease a digital set-top converter, the potential market for NVOD services in each node will be 350 customers. Assuming

the monthly movie buy rate among those customers is 100% (one movie, on average, per subscriber per month), the daily buy rate will be about 12. If one-third of those daily buys overlap in the busiest evening hour, then four simultaneous streams would be required. A single RF channel with 10-stream capacity would thus have 2.5 times the required capacity in 1/8 the spectrum required without switching.

A major technical issue with switched NVOD is reducing the time between subscriber channel selection and the appearance of video on the TV screen to the point that it is not noticeably slower than selection among nonswitched channels. Whether utilization of a switched architecture for delivery of NVOD and/or narrowcast channels is cost effective when compared with increased RF bandwidth is dependent on many factors, which vary with the system and will change with time as the cost of various technologies matures.

Note that the switched broadcast architecture shown in Figure 18.9 could be combined with VOD services in the same spectral range if some of the inputs to the MPEG MVX were fed by on-demand servers.

18.3 Architectural Examples

Engineering is the art of solving a technical problem for the lowest cost, given a well-defined set of requirements for performance, reliability, scalability, and other factors. Unfortunately for the engineer, the problem is not always well defined, though the budget for a given project may be. In that case, the task is to find innovative ways of meeting the short-term essential goals while planning for ill-defined future expansion. Such is almost always the case when designing the architecture for a cable television system. The many variants now deployed are testimony to the ingenuity of different designers. Thus, this section will concentrate more on the trade-offs among options than on comparing complete architectures.

The basic requirements are the same for any architecture.

- The end-of-line technical quality must meet corporate standards. At the least, it must meet FCC-defined performance levels with a sufficient margin to cover normal operating variations.
- The reliability and availability must be adequate for the services to be offered.
- The bandwidth must be sufficient to deliver all services to all customers. As a practical matter, the available bandwidth will be divided among commonly-delivered signals and those delivered to specific customers on a demand basis.
- The number of potential customers affected by any single network element failure must be limited to those who can be reasonably responded to by the customer service organization.

- The initial capital investment must be as small as possible, consistent with delivering the initial slate of services to the existing, or anticipated, initial number of customers.
- The network must be capable of cost-effective expansion to deliver additional services, to respond to higher utilization of initially defined services, and to respond to greater penetration among homes passed. This is especially true when the operator is not the incumbent operator with an established customer base. This requirement is known as *scalability*, and the best designs allow scalability that is specific to individual services and specific geographic areas.

18.3.1 Historic Development and Small-System Architectures

One of the earliest HFC architectures to develop was the cable area network (CAN), conceived as a rebuild strategy for older coaxial networks. The principal is illustrated in Figure 18.10. Fiber-optic links from the headend were used to

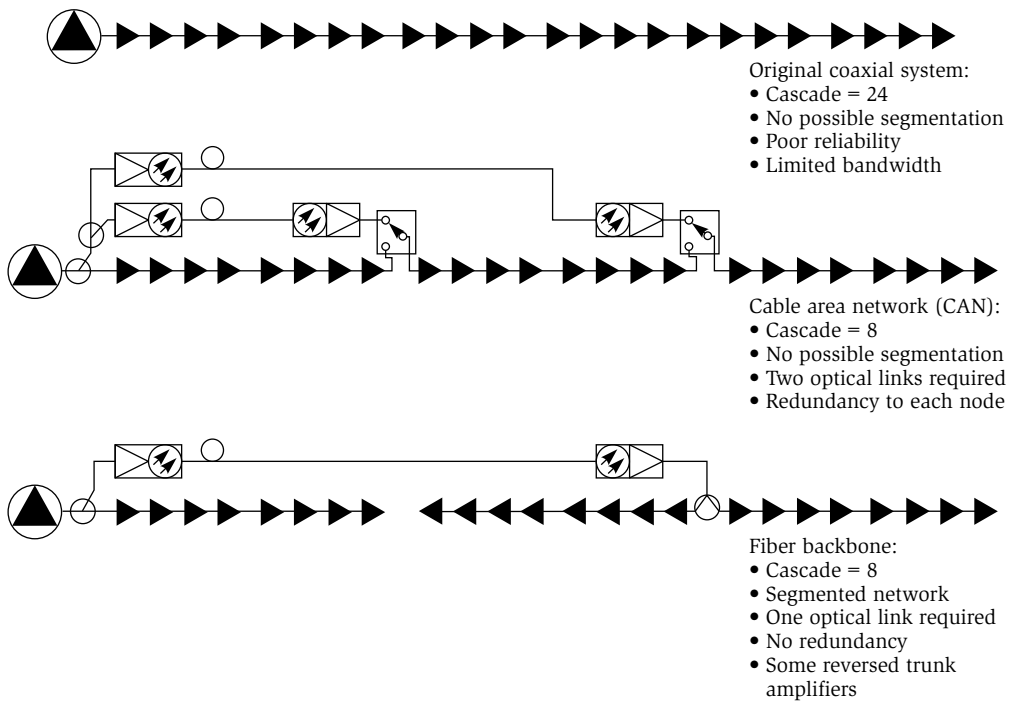


Figure 18.10 Principles of CAN and fiber backbone architectures.

transmit the common spectrum to selected nodes inserted at regular intervals in a former all-coaxial distribution network. This effectively shortened the cascades, which allowed the system to be upgraded to a greater RF bandwidth. Reliability, which was improved by shortening the cascade, was further enhanced by leaving the coaxial network between the nodes in place as a backup in case a fiber strand or transmitter failed.

Though a number of systems were upgraded using the CAN architecture, they were limited by being constrained to a common channel lineup. In fact, the architecture was useful primarily as a mechanism for minimum-cost bandwidth expansion and reliability improvement.

The first step toward truly innovative fiber architectures was similar to the CAN architecture, except that instead of just inserting nodes into an existing amplifier cascade, greater cascade reduction was realized by feeding both directions from the nodes. This was originally known as a *fiber backbone* and, like CAN, was proposed as a cost-effective system upgrade strategy. It became favored over CAN because, although it lost the redundancy feature, far fewer fiber links were needed for a given cascade reduction and fiber links were perceived to be highly reliable. The critical difference, however, was that the fiber backbone architecture had created the ability to independently program nodes, thus creating a simple star network.

18.3.2 Node Size and Architecture

A basic building block in any HFC architecture is the node, where optical signals are detected and launched into coaxial distribution cables, and the coaxial distribution subnetwork extending from there to customers. An important decision that fundamentally limits per-customer effective bandwidth is the number of potential customers served by each node.

As the cost of fiber cables, transmitters, and receivers has dropped and performance has improved, the smallest economic node sizes have dropped continuously. With sufficiently small nodes, it has become practical to offer bandwidth-intensive subscriber-specific services such as telephony, high-speed modems, and video on demand. RF bandwidths of 870 MHz combined with 500-home nodes are common for new construction, but many designs call for much smaller nodes, with a continuum from there down to fiber-to-the-curb (FTTC) designs that utilize individual drops to customers from the node. Also common are designs in which multiple coaxial distribution legs are fed from a common node location, with the node configured so that it can be divided electrically into separate logical nodes, one for each distribution leg.

Figure 18.11 shows one example of a three-step serving-area division plan that utilizes evolutionary capabilities built into some manufacturers' nodes. As initially installed, one optical receiver detects the commonly delivered analog signals and the other detects node-specific signals. After RF combining, the output is split to feed all four coaxial output legs, each of which is designed to

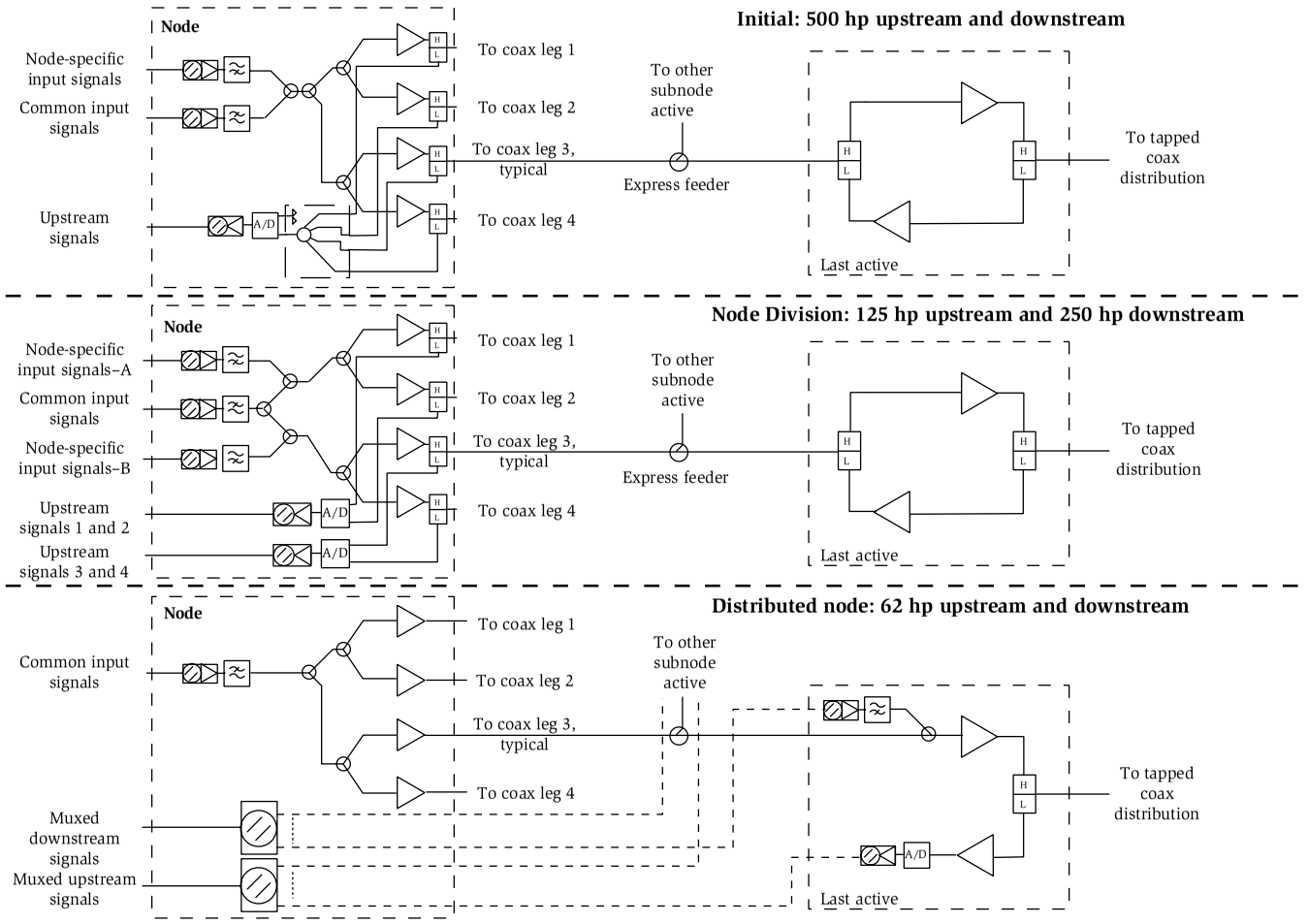


Figure 18.11 A typical node-scaling strategy.

deliver signals to approximately 125 homes. Each leg consists of an express feeder followed by an amplifier that drives tapped distribution cables. In the upstream direction, all four legs are combined at RF and fed to one input of a shared A/D converter, followed by an upstream transmitter. Thus, the node size is 500 homes in both directions. Three active fibers are required.

At the first level of scaling, a second node-specific downstream receiver is added and the downstream configuration changed to create two 250-home downstream virtual nodes, and a second two-input upstream converter/transmitter module is added (or a four-input A/D used — both are commercially available) to create four 125-home upstream virtual nodes. This asymmetry is consistent with predictions that upstream bandwidth will be saturated before downstream bandwidth. Five fibers are required to support this configuration, unless WDM is used. This is a very cost-effective upgrade, since only modules in the node, and the matching headend equipment, need be changed.

At the third level of evolution, fibers (dashed lines in the figure) are lashed over the express feeder to each amplifier location. Optical receivers are added there for only the node-specific programming, with the power and common signals traveling over the coaxial link. There are several options for upstream node division. In this example, CWDM transmitters are added at each amplifier location and a CWDM multiplexer at the original node location. On the assumption that eight amplifiers are ultimately required to serve the 500-home serving area, the final virtual node size has been reduced to approximately 62 homes. With downstream and upstream CWDM devices at the original node location, as few as three fibers from the hub would be sufficient to support this level of division.

18.3.3 Multilayer Networks

As pointed out earlier, pure star networks require a lot of glass and typically have single points of failure affecting many subscribers because the fibers feeding multiple nodes share the few available physical paths leading away from headends. These problems are proportionately worse as systems get bigger and as nodes get smaller. Given the trend toward regional consolidation and deeper fiber penetration, different architectures are required.

These larger networks include intermediate processing points called *hubs*. Depending on the specifics of the design, more than one intermediate layer may be specified. The addition of hubs offers some or all of the following advantages, depending on design:

- Shared use of fibers in the headend-to-hub link, thereby reducing the number of fibers required
- Provision of redundant transport between headend and hub, thereby improving reliability and reducing the possibility of outages affecting many subscribers simultaneously

- Use of intermediate signal processing at each hub to realize a net improvement in end-of-line quality
- Distribution of switched-service processing equipment to reduce instantaneous demand on headend-to-hub circuit capacity

There are two divergent opinions on how best to use hubs. One camp prefers to push as much signal-processing equipment out to the hubs as possible. This reduces the bandwidth requirement between headend and hub, because many signals are locally generated at the hub and only the multiplexed network-side signals need be transported back to the headend. For instance, VOD movies may be stored at hubs and streamed from there to specific customers on demand; only the occasional content updates need be sent down from the headend to hub-based servers. Similarly, telephone *host digital terminals* (HDTs) at hubs can terminate calls from subscribers, and only the multiplexed DSx signals are transported over baseband optical links back to the headend where the switch is located.

A second advantage is potentially improved analog video quality. If such signals are digitized, transported in that form to the hub, and reconverted to analog for transmission over the relatively short distance to customers, a higher end-to-end quality may be realized, especially where the total distance from headend to customer is at the outer limits of analog transport links. If the signals transported from headend to hub are all digital, the distance between them can be virtually unlimited, allowing arbitrarily large networks.

Offsetting these advantages, it is almost always more costly to divide processing among several locations than to do it all in the headend. Second, the more complex the signal processing at each hub, the more likely failures are to occur there. Then system operators are faced with a choice between a longer time to restore service outages and the higher staffing costs required to station personnel at each hub. Finally, the larger and more complex the hub, the higher will be the associated costs for real estate, emergency powering systems, air conditioning, etc.

The alternative multilayer philosophy centralizes as much signal processing as possible at the headend. Hubs are then used primarily or exclusively for optical processing and redundancy switching. For instance, DWDM transport may be used in a redundant sheath ring between headend and hub, with redundancy switching and DWDM demultiplexing there to separate the signals for each served node. Thus, the operator is able to realize reduced fiber usage and redundancy over the greater part of the path between the headend and the subscriber while utilizing a very compact and inexpensive hub. With 16-wavelength multiplexing it is entirely possible to build a hub serving more than 100 nodes in a half-dozen standard equipment racks, with similarly modest requirements for power and air conditioning.

Offsetting these advantages, purely optical nodes, though more efficient than pure star networks, typically require more fibers in the headend-to-hub link than hubs that include more local processing functions. Also, the use of

linear optical transport puts an upper limit on headend-to-hub distance and thus on system size. Although optical amplification is possible, each stage of amplification adds noise, and fiber-related cross modulation and second-order distortion increase with distance, as discussed in Chapter 13. Depending on the specific design, a typical upper limit is about 30–40 miles.

These architectural philosophies are not necessarily exclusive. A large system might well have several interconnected major hubs containing much of the signal generation and service-support equipment, each connected through linear optical links to smaller optical hubs that support clusters of nodes.

18.3.4 Distributed Architecture Example

Figure 18.12 is a simplified diagram showing a typical distributed architecture that illustrates the trade-offs possible. To keep the diagram uncluttered, the upstream paths from subscribers are not explicitly shown. It should be understood, however, that they terminate at the hub and are routed to the appropriate service equipment. Also, the optical transmitters feeding nodes are not shown. The distributed architecture is characterized as follows.

- All signals between the headend and hubs are transmitted in baseband, digital form, so distance limitations are removed.

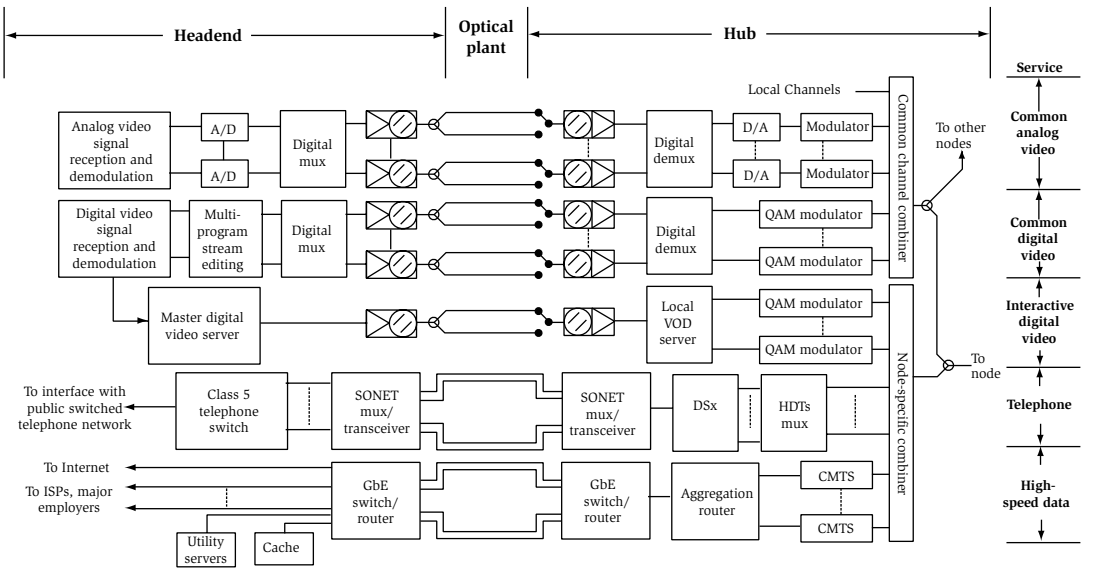


Figure 18.12 Headend–hub architecture with distributed processing.

- Each hub creates its own FDM spectrum for local nodes, thereby allowing the lineup to be customized on a regional basis (for example, for public, educational, and governmental access channels and regionalized advertising insertion.)
- Direct subscriber electronic interaction all takes place at hubs.
- Headend functions include "outside world" interfaces, master signal acquisition and manipulation, master content control for interactive video services, and network monitoring functions.

Within that general description, however, many variations are possible. For instance, the diagram shows diverse routing between two points, whereas in fact, a ring configuration is usually used that connects several hubs. In that case, some combination of repeating and tapped ring is employed, depending on the nature of the signals. Also, the diagram shows separate optical transceivers for each service category, though an operator may choose to partition the bandwidth of shared transceivers among several services. As an example, a SONET transceiver may be partitioned to carry both DSx services and gigabit Ethernet services. Finally, although a single master headend is shown, in fact multiple headends may be utilized for redundancy and/or because it is more efficient to centralize the equipment for specific services at different locations.

Analog Video Options

The form of digital transport for analog video channels is not shown in Figure 18.12. In order to avoid visible degradation, however, lossy compression techniques should be avoided. Options include high-bit-rate MPEG2 and proprietary constant-bit-rate schemes based on digitizing the signal at the common intermediate frequency of 41–47 MHz rather than at baseband.

Noninteractive Digital Video Options

The common digital video spectrum will almost always be delivered in multiprogram MPEG2 streams to set-top terminals. Multiprogram streams can be created at the headend or the hub, and the individual streams, whether multiprogram or individual, can be transmitted via MPEG transport, IP transport, DSx, or other common protocols.

Interactive Digital Video Options

The object of distributed processing for video-on-demand services is to stream individual programs to customers from the hub while managing the local server content from the central headend. As with noninteractive digital video, there are many options for sending content and control information to hubs, because it needn't be done in real time.

Telephone Options

In the example shown in Figure 18.12, the telephone service is provided over a proprietary constant-bit-rate (CBR) technology, using separate hardware and bandwidth dedicated to this service. With this option, host digital terminals (HDTs) provide the interface between the RF transmissions to and from subscribers and standard multiplexed telephone formats known collectively as *digital signals* (DSs). A hierarchy of DSs are available, including DS0 (a single digitized phone call), DS1 (1.544 Mb/s, comprising 24 DS0s plus overhead), and DS3 (about 45 Mb/s and comprising 28 DS3s). HDTs usually internally multiplex to the DS1 level, and an external DSx multiplexer combines those into DS3s, which can then be transported over SONET links between hub and headend, where they can be interfaced with a standard Class 5 telephone switch using a standard GR303 interface.

Should the operator offer telephone service using voice-over-IP (VoIP) technology, the telephone equipment at the hub would be eliminated. Instead, a router in the high-speed data equipment at the headend would sort voice from Internet packets and route the voice packets to a gateway device that would convert them to standard DS format for switch interface.

High-Speed Data Options

Each CMTS is capable of a downstream speed of 30–38 Mb/s. Therefore, the network-side interface needs to be capable of at least that rate. Depending on the design, however, multiple CMTSs may share a common chassis, with internal multiplexing to higher rates. In either case, various options are available for connection to an aggregation router, including 100BaseT and gigabit Ethernet. The aggregated hub data traffic may be transported, as suggested by Figure 18.12, by a gigabit Ethernet link but equally over a SONET link or any other link of adequate speed.

If VoIP services are provided over the shared data network, it is vitally important that the entire data network between the CMTS and voice switch interface respect the latency requirements necessary for transparent voice transmission.

Distribution Network Options

Figure 18.12 shows a simple combining of one leg from the common services combiner with the output of a node-specific combiner, which could be used to feed a dedicated transmitter for each node. Although an entirely usable scheme, most headends are much more complex. In general, each interactive service will be configured so that various numbers of nodes can share signals. Additionally, common signals and node-specific signals may travel over separate paths to the nodes, as shown in Figure 18.11. Various options for transmitting signals are discussed in Chapter 13; flexible node combining is discussed in the next section.

18.3.5 Centralized Architecture Example

Figure 18.13 is an example of the optical distribution network in a centralized architecture. As can be seen, it is characterized by:

- Separate processing of commonly distributed downstream signals (or, more accurately, area-specific, since some channels will generally vary from city to city in a metropolitan area), node-specific downstream signals, and upstream signals
- Generation of node-specific signals at the headend level
- Analog, QAM-modulated redundant transport of node-specific signals from headend to optical hub using DWDM to reduce the number of fibers required
- Analog, baseband-modulated redundant transport of upstream signals from hub to headend using DWDM
- An all-optical hub, with no RF or baseband signals except for status monitoring
- Flexibility in scaling of each interactive service within the headend

It is easiest to analyze this network in sections.

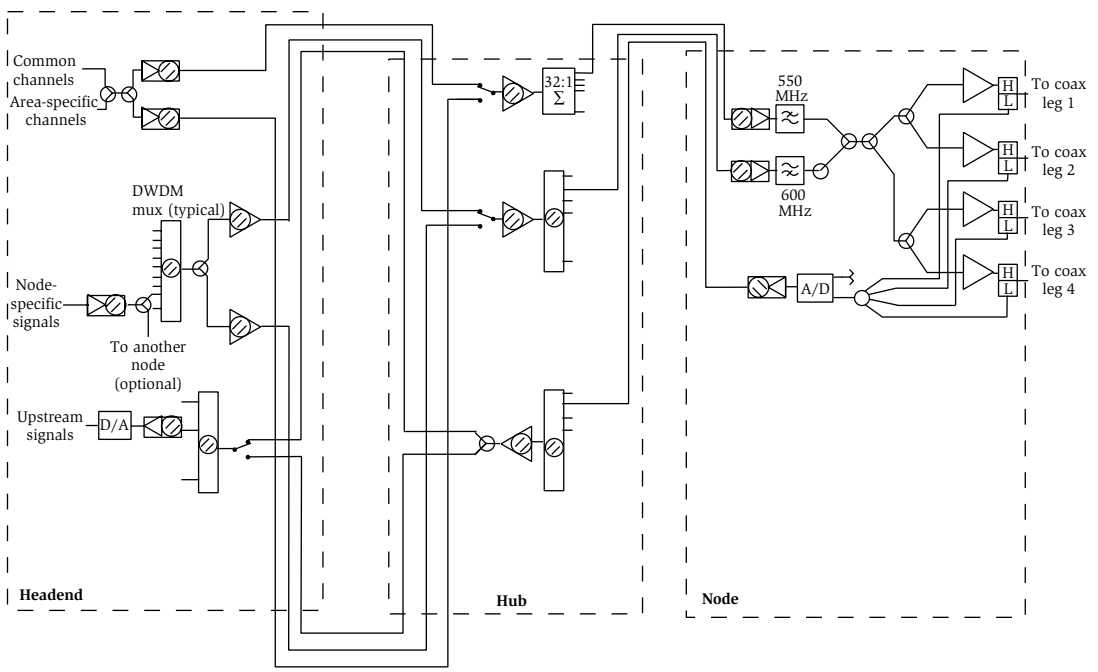


Figure 18.13 Optical transport portion of centralized headend–Hub–Node architecure.

Optical Transport Options

The optical transport architecture uses an externally modulated transmitter and a dedicated fiber for the analog-modulated video signals. At the hub, the signal is optically amplified and split 32 times, with one leg feeding each node on a dedicated fiber. Two headend transmitters drive separate fibers in the sheath ring configuration to provide redundancy against both transmitter and optical cable failure. DWDM is not an option for this link because the crosstalk mechanisms would result in interference below system specifications. Directly modulated transmitters are similarly not an option, because more than an octave is carried and self-phase modulation interacting with the fiber would result in serious CSO distortion.

The QAM signals directly modulate DFB transmitters that are combined using 16-wavelength DWDM between headend and hub. There, the combined signal is amplified and the wavelengths then separated, with a second dedicated fiber feeding a separate receiver in each node. Since less than an octave is carried and the output of the QAM detector is filtered before combining with the analog video spectrum, the second-order distortion is not a problem. Crosstalk mechanisms are a factor in determining the end-of-line C/N, but it is possible to achieve a 40-dB end-of-line system specification — adequate for 256 QAM signals.

There are other options that a system engineer should consider in deciding on an architecture for the downstream signals. Here are some examples.

- Common and node-specific signals could have been combined at the headend and used to modulate an externally modulated transmitter for each node. This would require fewer node receivers but more expensive transmitters as well as a dedicated fiber for each node.
- The common signals and node-specific signals could have been detected at the hub, filtered and combined, and then used to modulate a 1,310-nm DFB transmitter for transmission to the node, as shown in Figure 18.14. This would trade the cost of the optical amplifiers at the hub, the second downstream receiver at each node, and one downstream fiber for the cost of one DFB transmitter at the hub and the additional noise penalty from two series-connected transmitter–receiver pairs. It would also allow city-specific channels to be mixed at the hub level.
- The common and node-specific optical signals could have been combined optically at the hub, then transmitted over a single fiber to a single receiver at the node. As discussed in Chapter 13, however, this would restrict the node-specific bandwidth to less than 50 MHz to keep the second-order products below channel 2.
- The last option could be made workable with the full-bandwidth QAM signals if the headend-to-hub link were constructed with dispersion-shifted fiber, and very flat DWDM demuxes were used to reduce the second-order products.

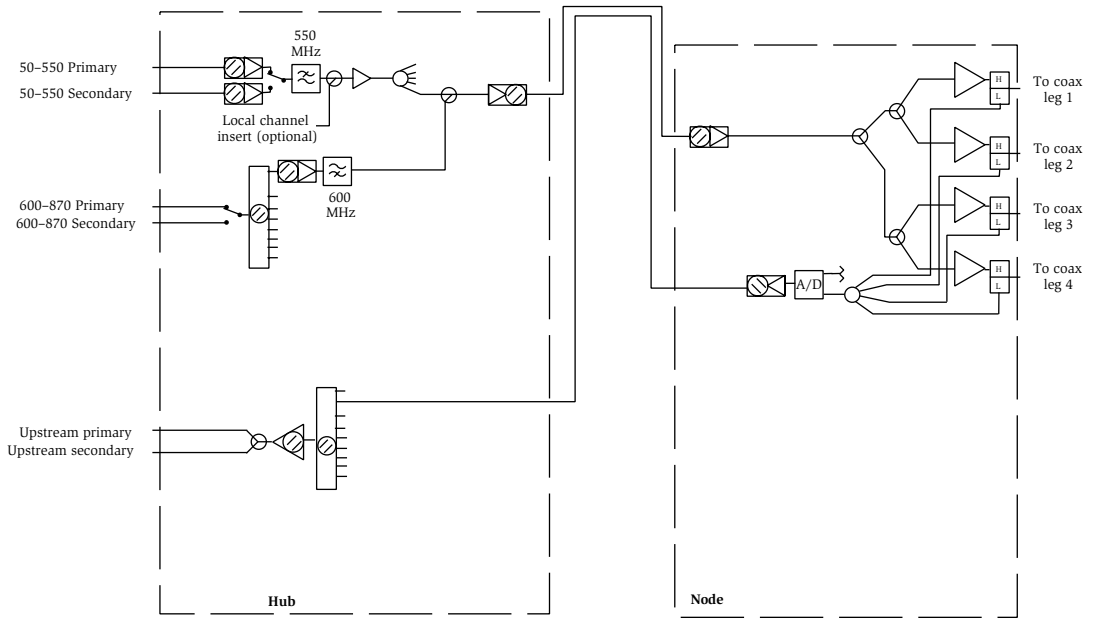


Figure 18.14 Optional repeating at hub.

- Although the upstream signals in the example were generated by converting the entire 5- to 41-MHz return band to a digital stream and transmitting as a high-speed baseband signal, the architecture would be similar if they were transmitted as analog signals. The trade-off is cost versus upstream optical budget and signal "durability" relative to degradation in the optical link.

There is no right answer here, just options that should be considered and evaluated against required performance, cost, scalability, and maintainability.

Optical Redundancy Options

Other choices determine the degree of redundancy. Figure 18.15, for instance, shows some alternatives for the analog common signal spectrum that could also be applied to the other signals:

- The *expanded redundancy option* shows that placing two optical amplifiers ahead of the redundancy switch would protect against failure of any active hub device for the cost of one high-power amplifier for every set of nodes.
- As the *highly redundant option* shows, using two optical amplifiers in the hub driving separate, diversely routed fibers to each node, with

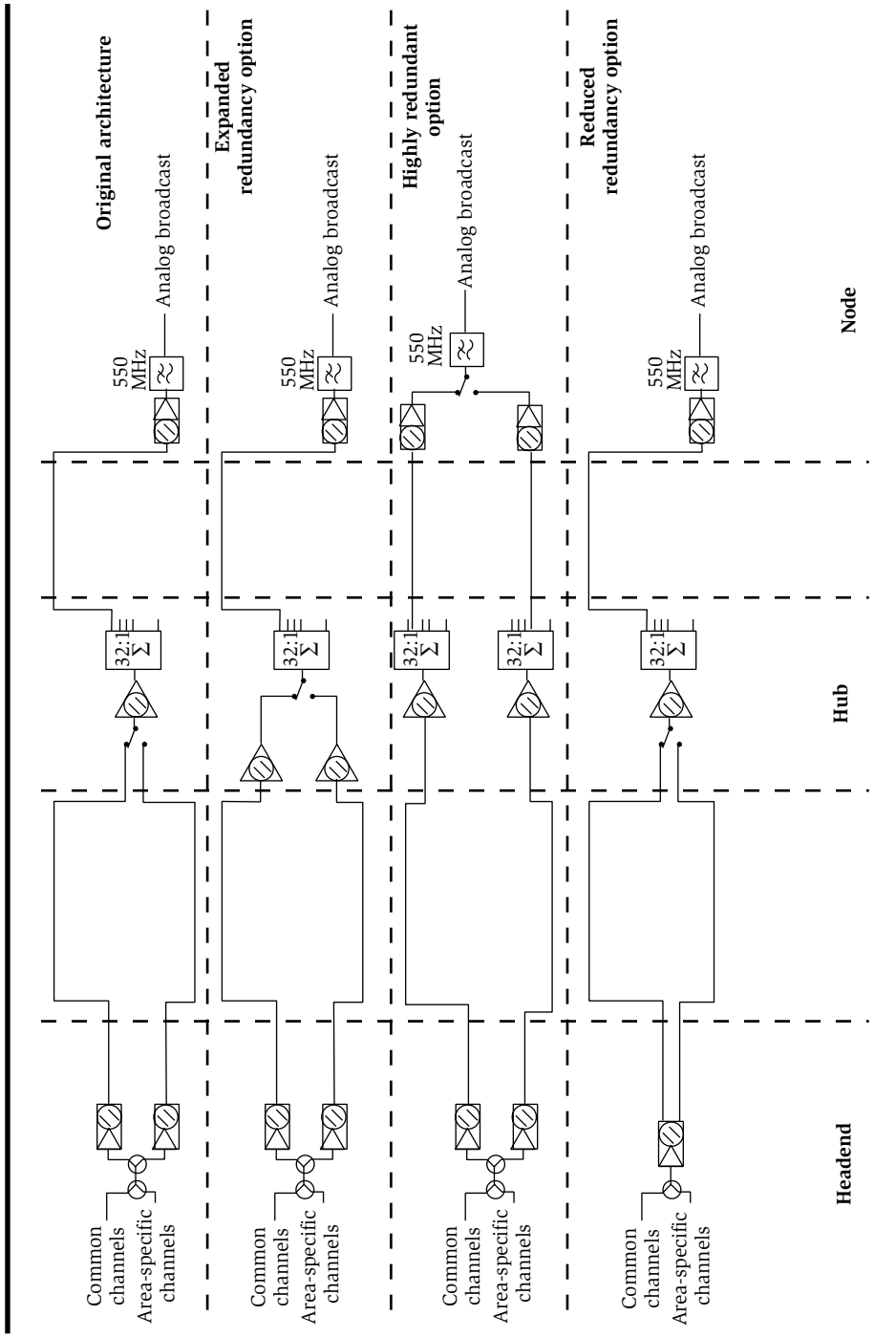


Figure 18.15 Examples of redundancy options for analog broadcast spectrum transport.

a redundancy switch after the detectors there, would protect against amplifier failure, cable cuts between hub and node, and node receiver failure. For highest redundancy, the two sets of hub equipment could be in separate facilities, offering added protection against a major facility failure (such as a fire). The use of redundant rings between hub and nodes is a feature of the Cox Communications "ring-in-ring" architecture.[6]

- Rather than two separate headend transmitters, a single transmitter can be used to drive both directions around the headend-to-hub ring, at a considerable savings in cost, as shown in the *reduced redundancy option*. In the case of externally modulated transmitters, this is particularly easy, since a second, isolated output is "free" because of the way Mach–Zehnder modulators work.

Scalability

Scalability is simply a way of expressing the ability of a system to meet growing demand for existing services or to deliver new services. What that means, on a technical level, is the ability to grow the service- and area-specific ratio of bandwidth to homes. Section 18.3.2 covered options for scaling the node and its serving area. Such scaling can obviously be applied on a node-by-node basis, as required, but it affects all services. Additionally, the downstream virtual node can initially be made larger by splitting the downstream QAM transmitters in the headend to feed multiple nodes.

Services can be scaled independent of the distribution plant scaling in a properly designed headend, however. As Figure 18.16 shows, each service-specific set of equipment is connected in an independent leg between the upstream service splitter and the downstream node combiner. One coax lead from each node's upstream applications splitter is wired to the input combiner/splitter network for each of the services shown, and one coax lead from the output combiner/splitter network for each service is routed to the node combiner area. The only exception is video-based services, where the input to each service (VOD in the example) comes from the shared set-top box (STB) transceivers. Within this general framework, two parameters can be varied to scale services in a node- and area-specific manner — the bandwidth allocated to the service and the number of nodes included in the service group.

Figure 18.17 shows an example of data scaling. Initially, the signals from four nodes are combined at each cable modem termination system (CMTS) input port. The equipment chosen has eight input ports, so the output is split to feed the same 32 nodes. A single 3.2-MHz, 16-QAM upstream RF channel and 6-MHz, 256-QAM downstream RF channel are used. The figure shows some of the evolutionary steps, which ultimately include elimination of upstream combining, addition of three more frequencies, and a reduction in the number of input ports used. In the ultimate configuration, this example allows upstream data

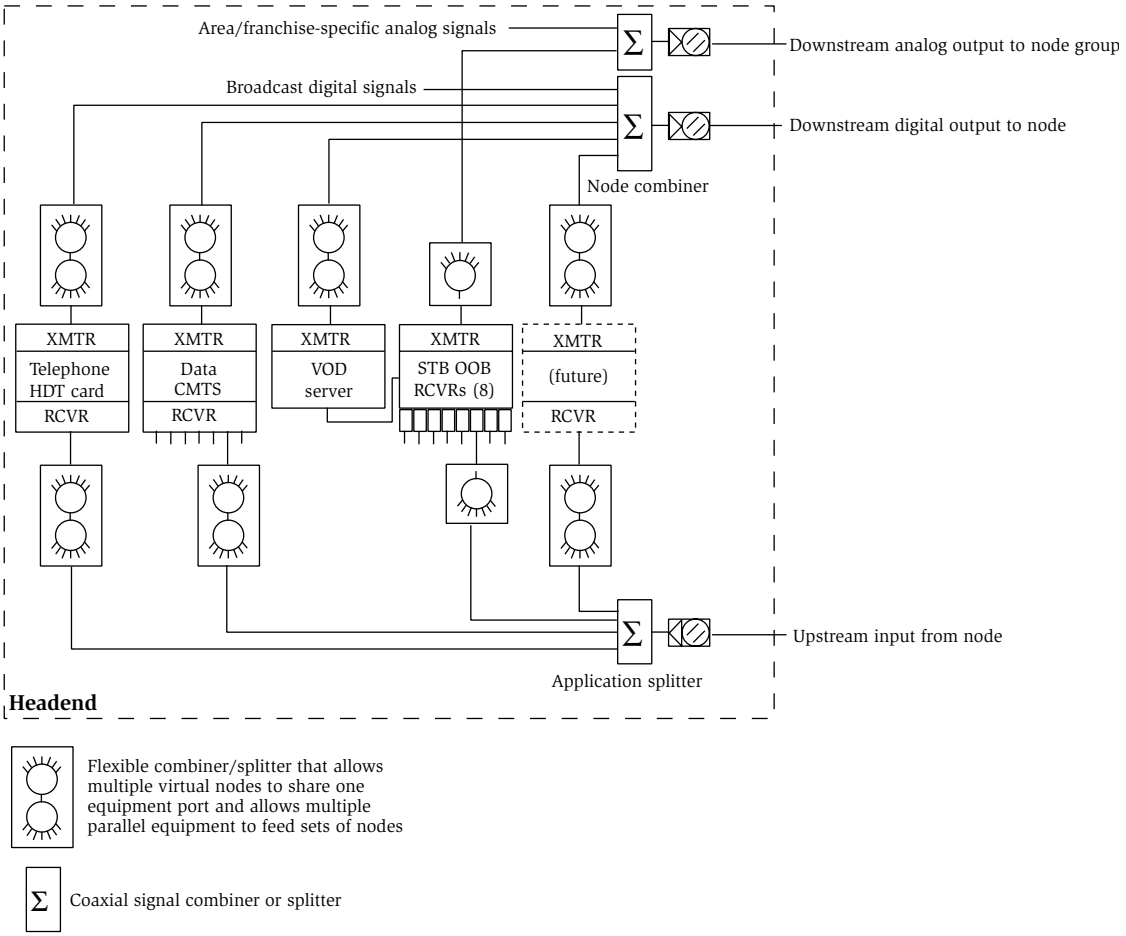


Figure 18.16 Flexible connection of service-specific equipment in headend.

scaling of 16:1 and downstream data scaling of 128:1, independent of any distribution system scaling.

Combining the distribution-scaling example of Section 18.3.2 with the first step in the preceding example of data-specific scaling would allow a total downstream data scalability of 256:1 and upstream data scalability of 64:1 within the same basic HFC architecture. Further assuming four 125-home coaxial legs served out of each node, a data service penetration of 50% of homes passed, and 20% simultaneous usage among subscribers, the system would be able to deliver average data rates of 6 Mb/s downstream to each simultaneous user and

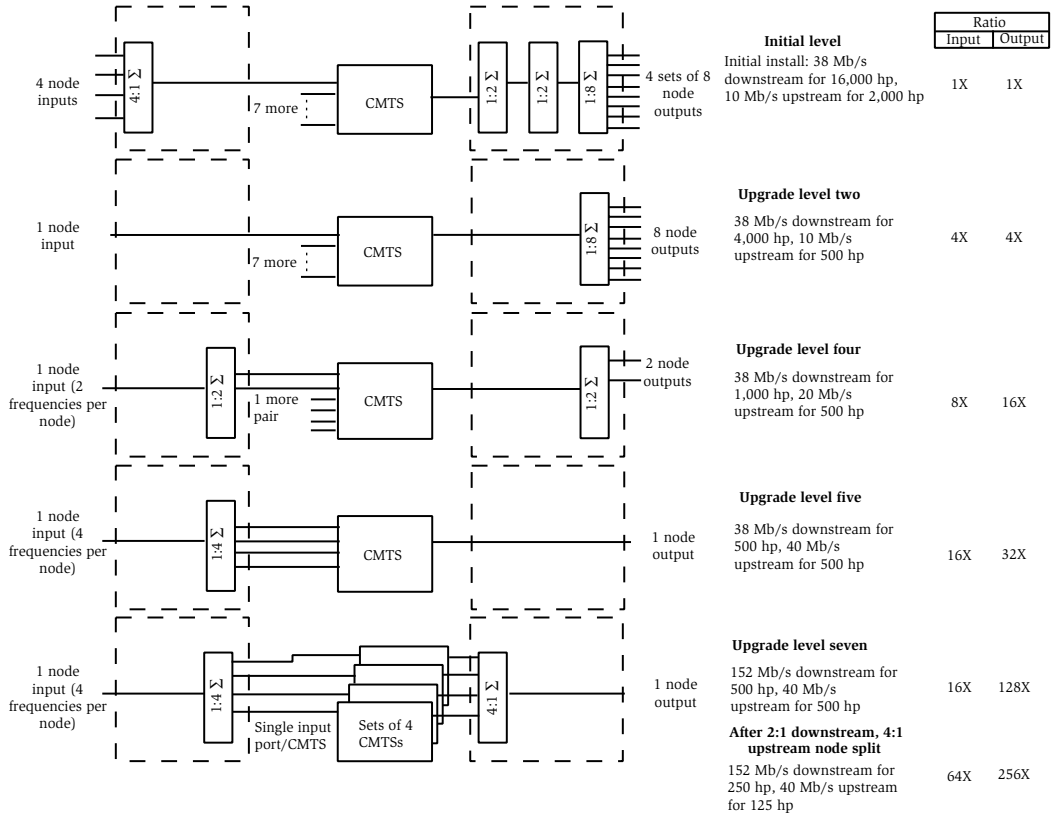


Figure 18.17 Headend high-speed data capacity evolution.

to receive 3.2 Mb/s upstream from each simultaneous user, all within the standard DOCSIS 1.0 or 1.1 framework and while utilizing only 24-MHz downstream and 12.8-MHz upstream bandwidth.

18.4 Summary

The architecture of a system determines what services it can deliver. It controls effective bandwidth, reliability, flexibility, and distribution of signal processing. This chapter has described some of the architectural elements and how each is related to essential network characteristics. The examples given illustrate choices made or proposed to meet specific service requirements in specific service areas.

The most common architecture in use today is the single star, with nodes connected directly to a single headend. In large regional systems, formerly independent headends are often linked by either digital or 1,550-nm analog fiber-optic links to a large master headend and become hubs. Single coaxial cable low-split-band plans are used in the vast majority of systems, with individual nodes serving 500–2,000 homes. Coaxial amplifier cascades vary from one to six in typical new upgrades.

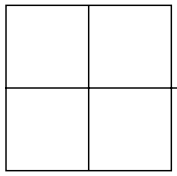
It is possible to "push" much signal processing out to hubs in order to increase the efficiency of the headend-to-hub links, or to centralize the processing for easier management at the expense of needing more fiber capacity. In the largest systems, both structures may be used in multitiered structures.

In considering an architecture, initial cost is certainly a factor. But just as important is the ability to scale to meet market demand and opportunities without "stranding" capital and without causing excessive service interruptions to existing customers because of required reconfiguration.

The relationship between architecture, service reliability, and service availability is a major topic in itself. This chapter has been limited to options within conventional HFC networks. Chapter 19 discusses FTTC and FTTH networks. The methodology for estimating reliability-related parameters is given in Chapter 20.

Endnotes

1. *National Electrical Safety Code*. Published by the Institute of Electrical and Electronics Engineers, New York, and approved by the American National Standards Institute as C2-1997.
2. Tom Osterman, Broadband Powering Issues, *Cable-Tec Expo '96*, Nashville, TN, June 1996, pp. 469–473. SCTE, Extor, PA.
3. Dan Kerr, Power Distribution in a Lifeline Network. *Communications Technology*, May 1997, pp. 64–70.
4. Leo Hoarty, Interactive Television — Architecture Under Test, *Symposium Record*. International Television Symposium, Montreux, Switzerland, June 1995, pp. 221–239.
5. Ran Oz, Amir Bassan-Eskenazi, and David Large, Unified Headend Technical Management of Digital Services. *1990 NCTA Technical Papers*, NCTA, May 2000, pp. 87–97.
6. Cox Communications Corporate Engineering Staff, Rings, Clamshells and Spurs. *CED Magazine*, February 1995, pp. 24–34.



Chapter 19

Digital Fiber Modulation and Deep Fiber Architectures

19.1 Introduction

This chapter covers binary (digital) modulation of a fiber-optic cable as well as deep fiber architectures. The two are covered together because binary optical transmission is used in last-mile (to the home and/or business) applications. In addition, binary optical transmission is used extensively in metropolitan loops and in intercity trunks. An advantage of digital modulation is that it can operate with much lower signal-to-noise ratio than can analog modulation, as shown in Figure 12.13.

Figure 19.1 illustrates the difference between digital and analog optical modulation. Figure 19.1(a) represents digital modulation of the optical transmitter. The data is not modulated onto an RF carrier. Rather, the data directly modulates the laser, turning it on and off. If more than one datastream is provided, they are *time division multiplexed* (TDM'ed) by switching first to one and then to another until all have been sampled. Since no data can be lost during the multiplexing process, the output data rate must be the sum of all input data rates, usually with additional bits added to synchronize data recovery. The data rate actually carried by the laser is called the *wire rate*.

Contrast this to the common cable television technique of broadcast optical transmission, as covered in Chapter 12 and shown in Figure 19.1(b). In this system, we often transmit digital signals, which may be TDM'ed but then they are modulated onto RF carriers, normally using either 64- or 256-QAM modulation. These digitally modulated carriers are combined with analog modulated carriers, and the sum of all signals modulates an analog laser, which operates in its linear range, where output power is proportional to input current. The use of multiple carriers to transmit different signals is called *frequency division multiplexing*, or FDM.

Usually the term *analog* is applied to the optical transmitter in Figure 19.1(b). It is true that this must be an analog transmitter, since a digital transmitter would never be able to carry the multiple frequencies without creating intolerable

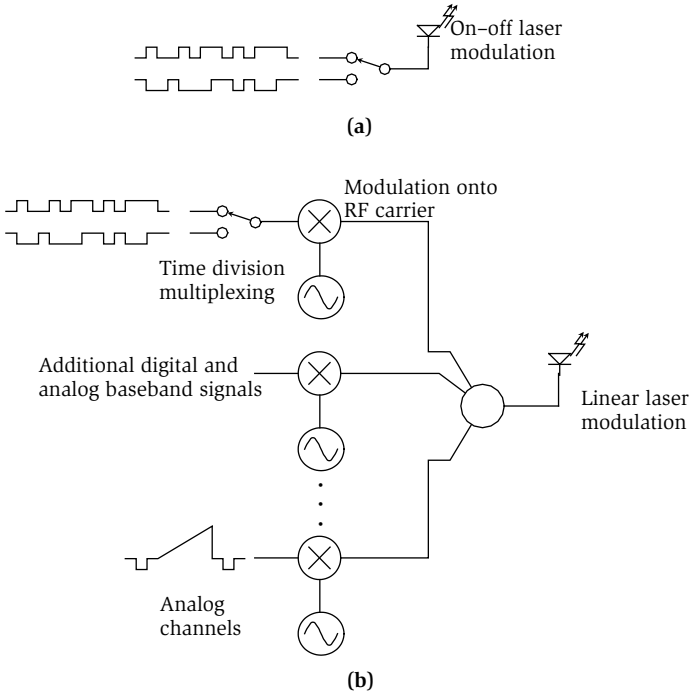


Figure 19.1 Difference between digital and analog optical modulation.

intermodulation distortion. However, the signals carried as modulated signals on the analog transmitter may themselves be either analog or digital.

19.2 Difference Between Analog and Digital Modulation

Figure 19.2 illustrates the difference between analog and digital optical modulation of a laser. Compare with Figure 12.13, which covers analog modulation only. The analog signal occupies primarily a range of diode currents in a more nearly linear range of laser operation. Note that the real analog waveform is not a sine wave; rather, it exhibits quite high peaks, as shown in Figure 12.20. Because of this, the carriers really exceed the range shown, with resulting clipping distortion, as described in Section 12.10.3. The levels must be managed so that the clipping distortion is not excessive. The digital signal is truly a rectangular waveform occupying but two levels, with fast transition between the two.

Analog transmitters are characterized by the optical modulation index, or OMI, as explained along with Figure 12.14. By contrast, digital modulation is

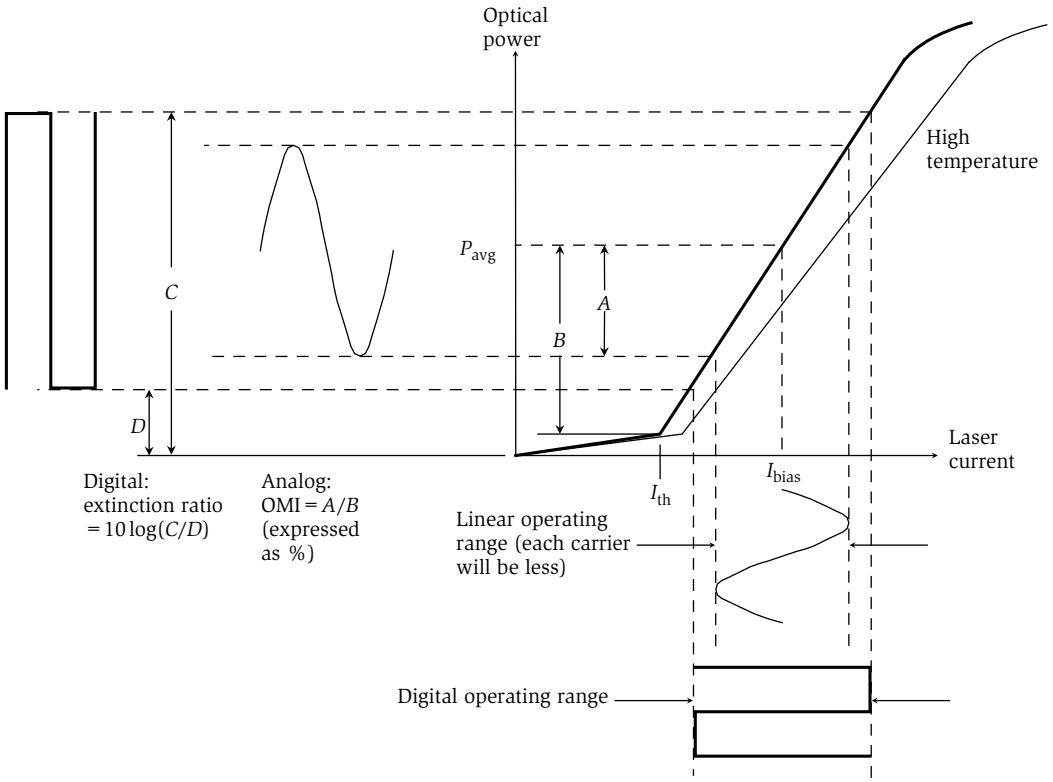


Figure 19.2 Modulation transfer functions for analog and digital optical modulators.

characterized by the *extinction ratio*, defined as the ratio between the maximum light output in one binary state to the minimum light output in the other binary state. Extinction ratio is expressed in decibels, as shown in Figure 19.2. The laser is biased so as not to extinguish the lasing process at any time, because restarting the lasing process causes problems in transmission. A high extinction ratio is desirable in order to maximize the ability to discern between the two states. However, over temperature and the life of the laser, the laser must never turn off during either of the two states. A typical minimum extinction ratio is about 9 dB.

19.3 Digital and Analog Transmitters

Figure 19.3 illustrates the block diagram of a digital transmitter in (a) and an externally modulated analog transmitter in (b). Data is supplied to the input of

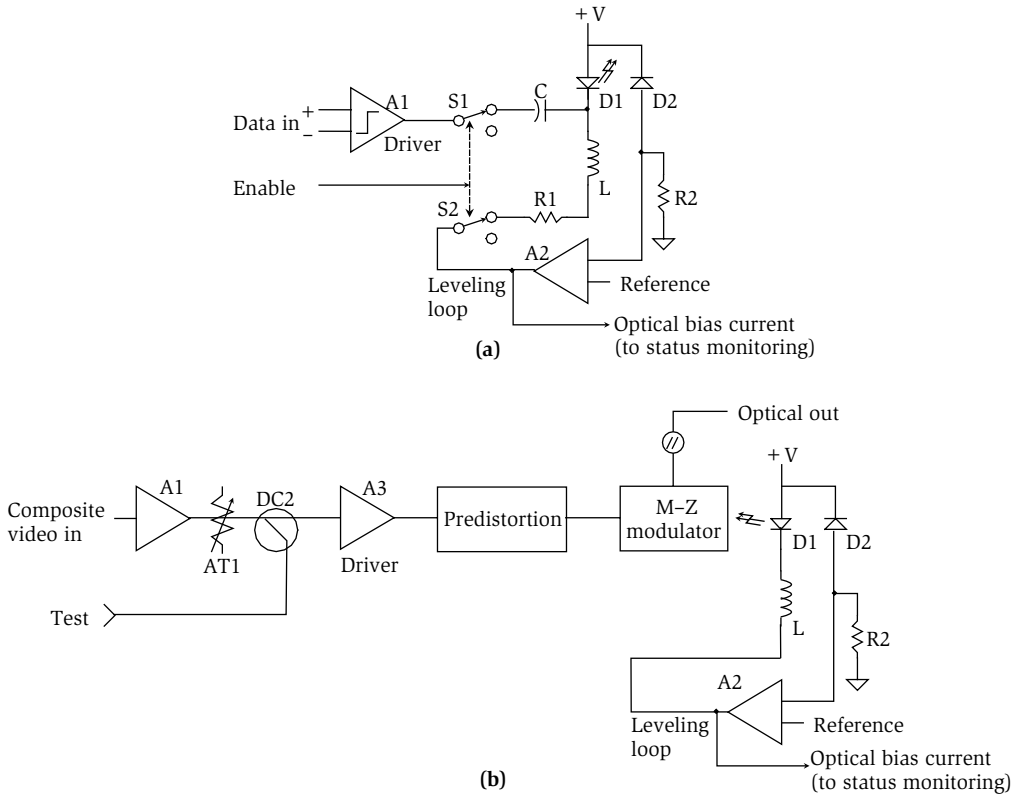


Figure 19.3 Optical transmitter block diagrams. (a) Digital. (b) Analog externally modulated.

the digital transmitter, usually balanced to minimize signal radiation and noise pickup. The driver amplifier A1 is a limiting amplifier, meaning that its output is not a function of the input amplitude. The drive signal to laser diode D1 is coupled through switch S1 and coupling capacitor C.

Monitor photodiode D2 samples the average light output from D1. The light is converted to a current in D2, and the current to a voltage in R2. The voltage appears on the inverting input of leveling loop amplifier A2, which compares the voltage with a reference. If the two differ, then the output of A2 changes, thus changing the bias in D1 and hence changing its optical output power. The output of the leveling loop amplifier A2 is often supplied to a status-monitoring (also known as element management) system, which monitors the current. The sensitivity of the laser diode decreases as it ages, resulting in more bias current needed to keep the light output constant. As the bias current approaches a specified threshold, it is time to replace the laser. This threshold may be

monitored in the element management system, and/or a comparator (not shown) may be added to the transmitter to generate a binary alarm signal.

Switches S1 and S2 are used when the transmitter must be switched off and on, as when it is used for time division multiple access (TDMA) applications. TDMA is explained in Section 4.3.1. When the laser is switched off, both the modulating signal from A1 and the bias from A2 must be removed to make sure the laser is extinguished.

Digital transmitters use F–P lasers for short distances and DFB lasers for longer distances. Because of their simpler circuitry and because the laser does not have to operate linearly, digital laser transmitters are significantly lower in cost than are analog transmitters.

Contrast this to the analog transmitter of Figure 19.3(b). For deep fiber applications, externally modulated transmitters are usually used because of their superior characteristics. Composite video is applied to amplifier A1. In a typical cable television architecture the composite input is the frequency division multiplexed set of signals that define the broadcast service. Attenuator AT1 sets the drive level. After driver amplifier A3, the signals are usually sent to a predistortion circuit, which compensates for distortion in the laser and signal path, by applying an opposite distortion to the signals before they are modulated onto the laser. Finally, signals are coupled to an external modulator, such as the Mach–Zender (M–Z) modulator described in Section 12.6.4. Monitor photodiode D2 serves to keep the output light level constant, as in the digital transmitter.

19.4 Digital and Analog Receivers

Figure 19.4 illustrates digital and analog receivers. In the digital receiver of Figure 19.4(a), detector diode D1 receives the light and converts it to a current. The average current is filtered in L1 and converted to a voltage in resistor R1. Signal output is developed across L1, and AC coupled through capacitor C1 to a *transimpedance amplifier* (TIA). The TIA has a high input impedance and a controlled output impedance, such as 50 ohms. It is a linear amplifier, so the output waveform is an accurate representation of the input light power. Automatic gain control (AGC) may or may not be used. The output of the TIA is filtered in low-pass filter FL1 to remove out-of-band noise and improve the waveform. Finally, the signal is shaped in limiter A3, which converts the analog levels to logic 1 and logic 0. Typically the output of A3 is balanced, again to keep switching currents out of the ground system. An activity detector detects the loss of either optical signal or digital modulation on the signal and can report this fact to a status-monitoring system.

Contrast the digital receiver just described with the analog receiver shown in Figure 19.4(b). Light is again converted to a current in D1, and the signal portion of the current is impedance transformed in T1 before being applied to preamplifier A1. The gain is adjusted in attenuator R2 under control of either an AGC loop

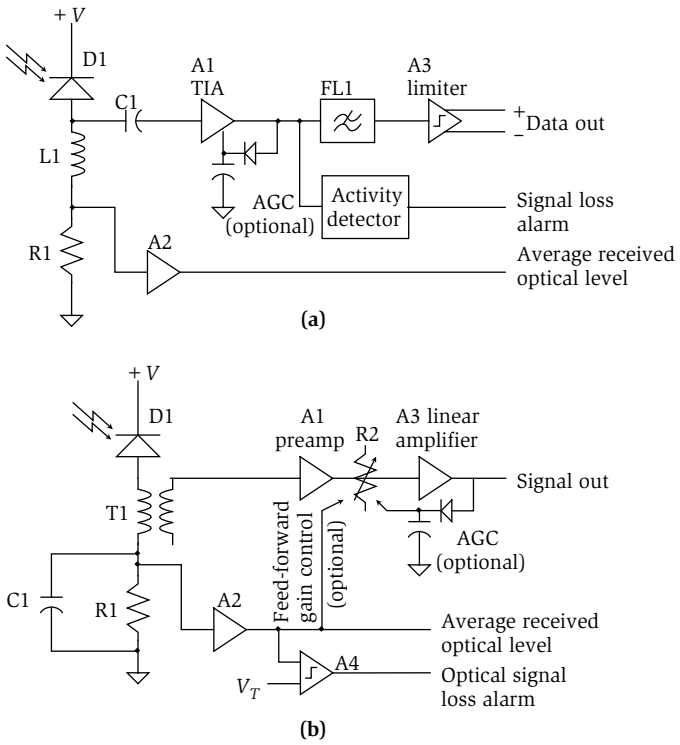


Figure 19.4 Optical receiver block diagrams. (a) Digital. (b) Analog.

or a feed-forward gain control circuit. A conventional AGC circuit is often used in optical nodes, but for other applications the feed-forward gain control may be used. The feed-forward circuit measures changes in the average optical level received and corrects the gain for the RF level changes produced by changes in optical level. This method is not as accurate as AGC, but it is lower in cost and is satisfactory for applications where level control is not critical. The average received signal level is derived as in the digital receiver, and a threshold detector A4 may be used to generate a loss of input alarm.

19.5 Combining Analog and Digital Transmission on the Same Fiber

It is quite possible to combine analog and digital transmission on the same fiber by using wave division multiplexing (WDM). Figure 19.5 illustrates WDM, in which one transmitter (usually the digital transmitter) operates at 1,310 nm, while the other (usually analog) operates in the 1,550-nm region. Other

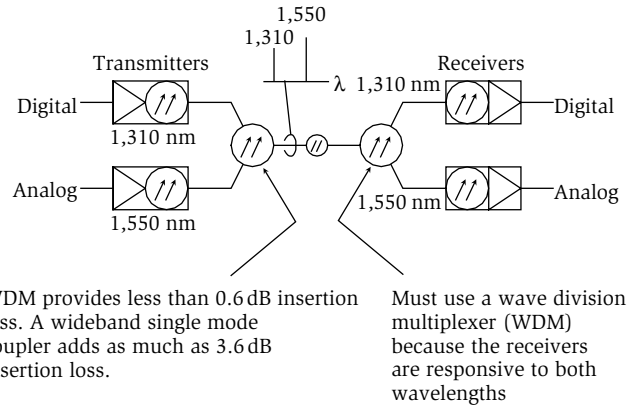


Figure 19.5 Combining digital and analog transmission on the same fiber.

wavelengths are sometimes used. Two or more wavelengths may be combined using a *wave division multiplexer*, which is the optical equivalent of a diplex filter. Alternatively, a single mode coupler, analogous to an RF combiner, can be used; but as with RF combiners, the signal will suffer insertion loss. Since receivers are responsive over a wide range of wavelengths, a wave division multiplexer must be used to separate the two wavelengths at the receiving end.

Wavelengths closer to each other, such as multiple wavelengths in the 1550-nm region, may be multiplexed. The ITU has specified standard wavelengths in this band. The technique of multiplexing closely spaced wavelengths is referred to as *dense wavelength division multiplexing* (DWDM), as discussed in Chapter 13.

19.6 Low-Frequency Content Removal in Digital Transmitters

As with most communications systems, a digital optical system is not able to transmit frequencies down to DC, so there must be some method of low frequency (sometimes called DC) removal. Note that the transmitters and receivers are both AC coupled. There must be a way of ensuring that the data does not contain energy below the cutoff frequency of the system. Besides causing problems with the AC coupling in the transmitter and receiver, content approaching DC implies that you can have a long time between state (1 or 0) changes. Clock signals are recovered by locking a local clock to state changes in the incoming data, so an excessively long time without a state change will allow the recovered clock to wander out of phase, damaging data recovery.

DC removal may be done in a number of ways. One method is to establish rules limiting the number of like symbols (1 or 0) that can be transmitted in a row. Another common method is to use a substitution set of symbols for each

Table 19.1 Example 8B/10B valid data code groups

Code Group Name	*Octet Value*	*Octet Bits* HGF EDCBA	*Current RD−* abcdei fghj	*Current RD+* abcdei fghj
D0.0	00	000 00000	100111 0100	011000 1011
D1.0	01	000 00001	011101 0100	100010 1011
D2.0	02	000 00010	101101 0100	010010 1011
D3.0	03	000 00011	110001 1011	110001 0100

byte of data to be transmitted. This technique is used in gigabit Ethernet transmission.[1] The substitution method used in gigabit Ethernet and in certain other applications is called 8B/10B encoding. For every 8 bits (one byte), a 10-bit symbol is substituted, hence the name 8B/10B encoding. The substituted 10-bit symbol is chosen to have very close to an equal number of 1s and 0s and three to eight transitions per symbol. The codes satisfy the requirement of no DC component in the signal, and the large number of transitions ensure clock synchronization. Furthermore, since a limited number of the available codes are used, the encoding provides another way to detect transmission errors. The penalty is that, since 10 bits must be transmitted to represent 8 bits, the bandwidth required is increased by 25%. For instance, in a gigabit Ethernet system, the desired data is transmitted at 1 Gb/s, but because of 8B/10B encoding the data rate on the fiber (the so-called wire rate) is 1.25 Gb/s.

Table 19.1 illustrates a few of the many valid code groups used in gigabit Ethernet systems. The two columns labeled "Current RD−" and "Current RD+" represent two so-called *running disparity* sets. The running disparity rules change the transmitted value from one column to the other based on certain rules related to the number of 1s or 0s that have been transmitted in the previous code group. These rules ensure that there is no DC content and that there is not a long string of like binary digits, thus ensuring reliable clock recovery.

19.7 Bidirectional Transmission

Almost all digital networks are bidirectional full duplex links, in which transmission takes place simultaneously in both directions. Full duplex transmission is accomplished differently in different situations. For long-distance transmission using either ATM or Ethernet, it is common to use different fibers for transmission in the two directions. Because of the high volume of transceivers for these applications, it is possible to buy low-cost, very small transmitters, sometimes called GBICs in the Ethernet world, specified for different distances to be covered.

Some systems intended for limited (perhaps up to 10 km) transmission distance transmit in both directions on the same fiber at the same wavelength

(usually 1,310 nm). This works for digital transmission because the signal-to-noise ratio required is substantially less than that required for analog transmission and also because the isolation in optical couplers can be quite high (greater than 55 dB). In order to make single fiber bidirectional transmission work, it is necessary to handle digital optical connections the same way you would handle an analog optical connection, in terms of cleaning and the use of angle polished connectors. The handling techniques required are described in Section 12.5.1.

Some systems use a third wavelength to transmit digital signals in one direction. For example, 1310 nm might be used for upstream digital transmission, 1490 nm might be used for downstream digital transmission, with 1550 nm reserved for analog transmission. This plan is used in the FSAN standard, ITU G.983.3 described in Section 19.9.3. Figure 12.6 shows that optical fiber can transmit light at these wavelengths with low attenuation. A special WDM element with three passbands at 1310, 1490, and 1550 nm is used to separate or combine the wavelengths with lower loss than is possible using a wideband coupler.

19.8 Fiber-Deep Architectures

Several architectures are being used to drive fiber very deep into the network, even as far as the home. These architectures are driven by the improved quality of signals delivered over fiber-optic plant as compared to those delivered over coaxial plant, by improved reliability due to fewer devices in the network, and by improved bandwidth. The lack of amplifiers can mean lower maintenance costs due to the lack of a need to sweep and balance amplifiers. At least in FTTH systems shown in Section 19.8.2, the RF signal level never approaches 10^{-4} W (+ 38.75 dBmV), the level above which composite leakage index (CLI) measurements must be made. Thus you can avoid the cost of making measurements and filing reports.

19.8.1 Fiber-to-the-Curb Architectures

In *fiber-to-the-curb* (FTTC) architectures, fiber-optic cable brings signals to a termination point, where they are converted to electrical form for distribution to a small number of subscribers. From the termination point, coaxial cable is used for video services, telephone cable is used for telephone service, and Cat5 or Cat6 cable (see Chapter 5) is used for data. Characteristically, no amplification of broadcast signals is provided after the optical termination point in systems that distribute broadcast signals. Not all FTTC systems have included distribution of broadcast signals.

Figure 19.6 illustrates a typical FTTC system. One or more optical fibers extend from the headend to an optical termination point. Besides the normal conversion from optical to RF for the broadcast information (the downstream

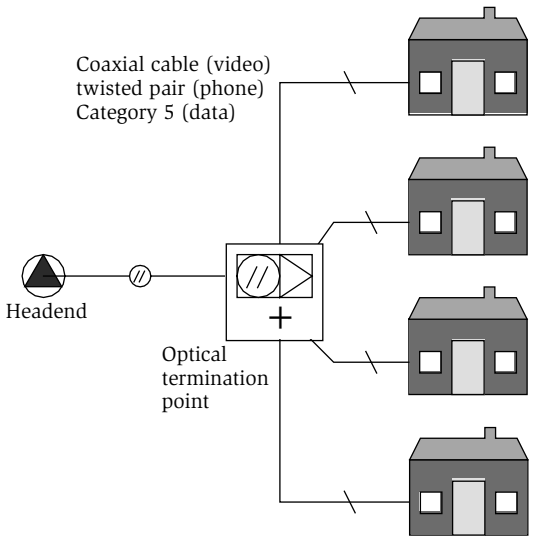


Figure 19.6 Typical FTTC architecture.

function of an HFC node), the optical termination point includes a bidirectional digital optical interface and interfaces for telephone and data networks. These are covered more fully next (fiber-to-the-home systems).

19.8.2 Fiber-to-the-Home Systems

Fiber-to-the-home (FTTH) systems have been built in several different configurations. They can be subdivided into several types, but all have in common that they extend to the home the interfaces found in FTTC systems.

Passive Optical Networks (PONs)

Figure 19.7 illustrates a basic FTTH system using a *passive optical network* (PON). Terminology borrowed from the telephone industry is often used in such systems to define the endpoints. The headend (or hub or central office) side is known as an *optical line termination* (OLT). For a PON, this would include the broadcast optics and the routers used for digital transmission. The subscriber-side circuitry, which typically goes on the outside of the home, is called an *optical network termination* (ONT). Usually, the subscriber-side interfaces at the ONT include a broadcast coaxial connection (when broadcast service is supported), one or more Ethernet connections, and one or more analog telephone interfaces.

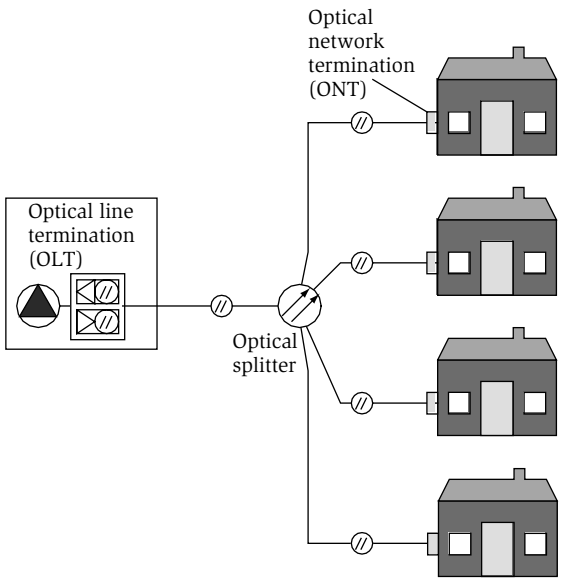


Figure 19.7 FTTH using a passive optical network.

One point of passive optical splitting is shown in the figure, but in practice two optical splitting points are often used. The primary advantage of a PON over active architectures described next is that there is no active equipment in the field between the hub and the subscriber. No power is required in the field. A disadvantage is the limited distance and a limited number of times the signal can be split, as shown next. Furthermore, more expensive optics may be required for high-speed data.

PONs are differentiated by the underlying layer 2 transport system used (see Chapter 5). Some are based on ATM and are known as APONs. Others are based on Ethernet and are known as EPONs. All deliver data to the home, with Ethernet connections being the dominant subscriber-side connection. In the case of APONs, ATM is transparent to the subscriber.

If the PON carries analog (frequency division multiplex, using both analog and digital modulation) signals for broadcast program delivery, some people add the designation BPON, for broadcast PON. Not all PONs carry analog signals. Some are intended solely as data networks, and some assume that video will be carried only in IP packets (IPTV).

Active FTTH

Figure 19.8 illustrates an active FTTH system. It is differentiated from a PON by the addition of active processing equipment in the network. Power is required at

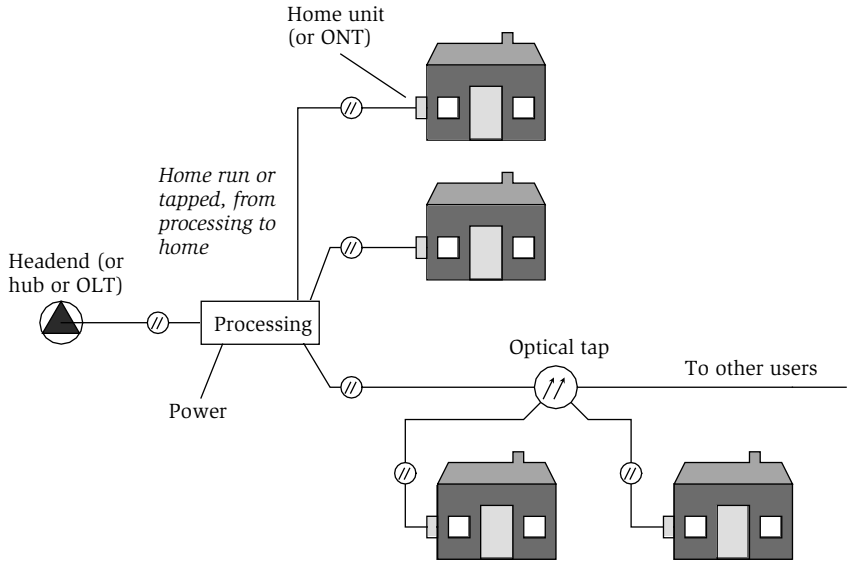


Figure 19.8 Active FTTH system.

that equipment. In known systems, there is only one active component between the hub and the subscriber. Beyond the processing point, fiber may be home run (a dedicated fiber or fibers exit the active equipment for every home), or a fiber may be shared using optical taps, analogous to RF taps. The active equipment may be installed in a roadside cabinet, a controlled environmental vault (CEV), or may be strand or pedestal mounted, as is an optical node or an amplifier in an HFC network.

The disadvantage of an active fiber system is that power must be provided for the active point, though the system may consume less power per home than do HFC networks. The advantages of having the active device include a much greater reach from the hub to the home and greater degrees of splitting. A further advantage is lower costs due to the ability to employ lower-cost optics at the home. The active system also enables network powering of the home unit if desired. Otherwise, the home unit is powered from the home, using a battery-backed power supply if desired.

19.8.3 Classifying Fiber-Deep Systems

Figure 19.9 is a classification of the different fiber-deep systems. We shall classify them first as either fiber-to-the-curb (FTTC) or fiber-to-the-home (FTTH) systems. FTTC systems are often passive from a neighborhood optical termination

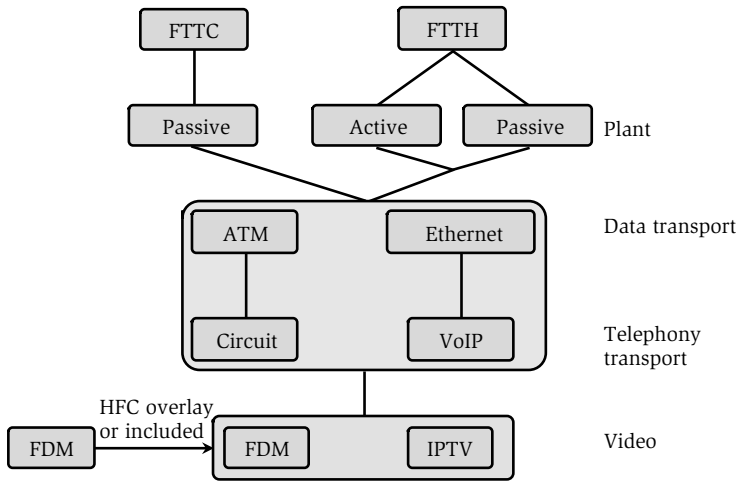


Figure 19.9 Classification of fiber-deep systems.

point (OLT), but the OLT is active and can be rather large for the number of subscribers served. FTTH systems are more prevalent and may be all-passive or may have some active components, as shown in Figure 19.8.

Data transport on the network may be based on either ATM or Ethernet (see Chapter 5). ATM has its background in the telephone industry and is known for excellent quality of service (QoS) capability. Ethernet has its background in local area networks. Traditionally Ethernet has not incorporated QoS features as are available in ATM, though recent advances have allowed Ethernet to take on QoS features similar to those of ATM.

Telephony transport tends to be circuit-switched-based in ATM-based systems and VoIP-based in Ethernet-based systems. Chapter 6 provides more information on these two options.

Finally, video delivery is usually traditional broadcast. We have used the term *FDM* for this, since the signals are modulated onto many RF carriers and transported over linear networks. Both analog and digital modulation can be, and are, used. A variation is one system that uses an HFC overlay for video and an FTTH system for data. Where an existing HFC network is in place and provides adequate video service, you may want to start with an overlay for data, emphasizing business applications, and later migrate video delivery onto that system, eventually phasing out the HFC plant at the end of its life.

The alternative video delivery mechanism is to send video via Internet Protocol as part of the data stream (IPTV — Internet Protocol television). Given adequate bandwidth, this works. Some providers use IPTV exclusively, but others see it as an adjunct to the FDM delivery of most video. An interesting

concept is to consider sending video on demand (VOD) over IPTV while using FDM for all other video services. Since VOD is by definition a point-to-point service being used by only one subscriber, this may be a better use of bandwidth than tying up broadcast spectrum when it is serving only one subscriber.

19.9 Distance Limitations in Fiber-Deep Systems

Fiber-deep architectures must trade off distance and the number of times the signal can be split. Each time the signal is split, loss is introduced, as described earlier. There are practical limitations on the amount of power that is injected into the fiber, and there are minimum optical signal levels that are required at the receiver to yield good performance. One distance limitation is the *tolerable loss* in the fiber network (defined as the difference between injected signal and minimum satisfactory received signal level). Sometimes the distance may be limited by other criteria. A budget may be constructed for various cases.

19.9.1 Limitations on the Distance of Analog Transmission

Stimulated Brillouin Scattering (SBS)

SBS limits the maximum power that can be injected into the fiber. When the power level of monochromatic light injected into a long fiber strand is increased, the output power increases proportionately until a threshold is reached. Beyond that level, the received power stays relatively constant, and light scattered back toward the source increases dramatically. The carrier-to-noise and signal-to-distortion ratios of the received signals both degrade. This phenomenon, known as *stimulated Brillouin scattering* (SBS), is described in Section 12.4.6. SBS limits the maximum power that can be injected into the fiber. As the state of the art in optical components has improved, the maximum injection level has increased. The current state of the art is about +16 dBm optical power injection into a long fiber at any one wavelength. This high power is achieved by applying any of several techniques to spread the spectrum of the emitted light. The reason spreading the spectrum works is that SBS is a function of light energy at one wavelength. Transmitting at multiple wavelengths partially mitigates the effect of SBS but introduces dispersion issues.

Shot Noise

A significant limitation in the performance of analog optical receivers is shot noise. Shot noise is generated in the receiver photo diode, caused by the statistical variation in the arriving photon distribution. The subject is covered in Section 12.8. As the modulation of the light source (the OMI) increases, C/N_{SHOT} also

improves. Higher diode responsivity also improves the carrier-to-shot-noise ratio. When shot noise is a significant contributor to total optical link noise, the total C/N (carrier-to-noise ratio) will change less as a function of optical receive level than would be the case based on just thermal noise considerations.

Thermal Noise in the Fiber-Optic Receiver

The heart of an optical receiver is the photodiode, which has an output current proportional to the light power impinging on the diode. The photodiode, being a current output device, is often coupled to a transimpedance amplifier (TIA), which amplifies the current and converts it to a standard impedance, usually either 50 or 75 ohms. The C/N of the output signal is dependent on the shot noise current in the diode, as shown earlier, and by thermal noise in the amplifier(s) following the photodiode.

As a practical matter, most optical receivers designed to receive analog carriers are limited to a minimum light input level of about −5 dBm. For node operation in HFC plants, the minimum level will be somewhat higher. This is because a portion of the noise budget must be reserved for coaxial amplifiers following the node in HFC networks, whereas in most fiber-deep systems there is no amplification following the node.

Optical Amplifier Noise Contribution

Section 12.7 describes the most common type of optical amplifier, an EDFA. Its noise contribution is the final significant contribution to the noise performance of an analog link. The noise added depends on the input signal level and on the amplifier's noise figure. When several amplifiers are used in cascade, their noise contributions add, just as with RF amplifiers.

Other Noise Contributions

There are a few other noise contributors besides the ones just described, which as a practical matter tend to be of a less limiting nature. These include laser relative intensity noise, which is defined in Section 12.6.

Analog Transmission Distance Limitation

We have established that the maximum optical signal level into the fiber is about +16 dBm, limited by SBS, and that the minimum level at the optical receiver is −5 dBm. This yields a total loss budget for an analog link of 22 dB. We can convert this to distance, given various splits in the fiber. Optical splices have a loss of 0.1–0.2 dB, and connectors can have loss of upto 0.5 dB. We shall allow 2 dB for splice and connector loss, though more may be necessary in the real

Table 19.2 Approximate maximum distance versus split for analog transmission

Split	*Nominal Splitting Loss*	*Available Fiber Loss*	*Nominal Distance (1,550 nm)*
4 way	7 dB	10 dB	40 km
8 way	10.5 dB	6.5 dB	26 km
16 way	14 dB	3 dB	12 km
32 way	17.5 dB	1.5 dB	6 km
64 way	21 dB	−2 dB	−10 km

See text for assumptions

world. This gets us to a loss budget of 20 dB. From this we can construct a table of the number of splits and the distance we can go without amplification. Note that this is based on nominal loss of the splitters. Real splitters may have more loss, shortening the distance.

Table 19.2 summarizes the maximum distance you can transport analog signals, given no amplification in the plant and using state-of-the-art components. We have started with the loss budget, subtracted connector loss, subtracted the nominal splitting loss, and allocated the remainder to fiber loss, converting this to distance. This assumes we start with the highest power we can supply to the fiber without encountering problems with SBS. We also assume high OMI in the transmitter (on the order of 3–3.5% per carrier), required to achieve good C/N. This table is just an estimate, based on thermal noise being the limiting factor in length. The table is constructed so that with modern components you can achieve a C/N of 48–49 dB. Some practitioners have gone to as few as a four-way split to make sure they have adequate range, but with splits this small PONs force a very large number of optical cables and transmitters at the hub. Notice that we have included the maximum distance for a 64-way split, tongue-in-cheek. This split has so much loss that you cannot make it work for video at all, given the best components available today.

Of course, optical amplification may be used to achieve longer distances. This is done in active FTTH systems. Adding an amplifier converts a PON into an active system.

19.9.2 Limitations on the Distance of Digital Transmission

Signal-to-Noise Ratio

As a practical matter, digital transmission, where the laser is turned on and off (amplitude shift keying, ASK), is limited by thermal noise, mode partition noise, and crosstalk. Figure 12.12 shows that the signal-to-noise-density ratio required

for digital transmission is significantly lower than that required for analog transmission. The receiver includes a low-pass filter that limits the noise bandwidth. When the bandwidth of the transmission is doubled, the low-pass filter cutoff frequency must also double, meaning that the noise power doubles. Thus, every time the data rate doubles, the received signal power must also double. You can double the received power by halving the number of times the signal is split, by reducing the distance drastically or doubling the transmitted power.

Some PON architectures have reduced the upstream data rate to permit lower-powered transmitters at homes, and some have limited the number of splits for the same reason. The object is to use lower-power, low-cost transmitters at subscribers' homes.

Mode-Partition Noise

Particularly in the upstream direction but also in the downstream direction, it is important to use low-cost lasers. The lowest-cost lasers in common use today are F–P (Fabry–Perot) lasers, though VCSELs are showing great promise. F–P lasers are multimode lasers, which produce light over a band of wavelengths. The amount of power at each wavelength tends to vary randomly due to the reflective structure in the laser. Without modulation, the total power is reasonably constant, but the wavelength at which that power is emitted changes. This phenomenon is known as *mode instability*. By itself, mode instability is harmless, but when mode instability exists and the light passes through a fiber-optic cable having dispersion, the result is *mode-partition noise* (MPN), a limiting factor in transmission distance. Dispersion is discussed in Section 12.4.3, and MPN is discussed in Section 12.6.2. Dispersion is the characteristic of optical fiber by which the velocity of propagation is a function of the wavelength of the light.

As the light output from the laser shifts between wavelengths, the velocity at which the pulses propagate changes due to chromatic dispersion in the fiber. This means that pulses recovered at the receiver will be "smeared" in time. If the smearing becomes excessive, the pulse cannot be recovered, and the bit error rate (BER) increases. A frequent measure of the effect is to determine the amount by which received power must be increased in order to get back to the same BER as would obtain without MPN. A useful relationship is the inequality

$$\sigma\text{BLD} = \sigma \times \text{bit rate} \times \text{distance} \times \text{dispersion} < 0.15$$

where σBLD is an acronym for bit rate, distance (L), dispersion. The laser RMS linewidth is σ. Dispersion versus wavelength — chromatic dispersion — is described in Section 12.4.3. The inequality is dimensionless when compatible units are used for all parameters. This relation ensures that the power penalty is low. In order to achieve longer distance, you can reduce the bit rate (this is going in the wrong direction), or you can decrease the linewidth of the laser (that is, the spectrum over which the laser spreads its energy due to MPN) by choosing a

suitable (more expensive) laser. Reducing the dispersion in the fiber is rarely practical. The dispersion issue involves the uncertainty of fiber zero-dispersion wavelength (1,300–1,324 nm) and the laser wavelength (including unit-to-unit variation and variation over temperature). Finally, you could shorten the distance (contrary to the idea of extending the fiber farther into the system).

Chirp

The term *chirp* comes from the way many birds make their chirps: Over the period of one chirp, the frequency of the pitch changes, going either up or down. In fiber optics, chirp is a change in the wavelength of the light. A directly modulated laser will change its output wavelength with changes in the laser current. This change in wavelength interacts with the chromatic dispersion characteristic of the fiber to produce deleterious effects, noted shortly.

Choice of Lasers

Except for very long distances, digital transmission can use lower-cost Fabry–Perot lasers. Due to their construction, F–P lasers tend to exhibit MPN. The total power from the laser is constant, but the wavelength of that power shifts. So long as the fiber-optic cable used exhibits near-zero dispersion at the wavelength of operation, this does not present a problem. Standard (non-dispersion-shifted) fiber exhibits zero dispersion around 1,310 nm. Without signal splitting, low-cost F–P lasers may be used up to about 10 km (1.25-Gb/s wire rate, or 1-Gb/s data rate in gigabit Ethernet systems). Higher-powered F–P lasers are capable of covering distances up to 40 km; above that, distributed feedback (DFB) lasers can be used.

Crosstalk

Crosstalk can occur in optical systems due to several mechanisms. In wavelength division multiplex (WDM) systems, the same fiber is used to transmit more than one wavelength. Several technologies are available to build components to separate the wavelengths, but all have loss, adding to the loss budget of the link. In addition, the attenuation of the undesired wavelength is limited, so there is a possibility of interference between wavelengths at the receiver. The interference also extracts a penalty in the minimum received power that will provide adequate BER. This is one form of crosstalk that can limit distance.

Another form of crosstalk occurs when the same wavelength is used for both upstream and downstream communications on the same cable. It is possible to separate signals flowing in the two directions, and the isolation between them is typically better than that obtained in RF transmission. However, it is not infinite, so there can be some interference between the near-end transmitted and received signals, again reducing the sensitivity of the receiver. In short-distance

transmission, there is frequently adequate margin and cost advantages to favor the use of simultaneous transmission in both directions on one fiber strand. In longer-distance transmission, two fibers are usually used to eliminate this cross-talk penalty.

Digital Transmission Distance Limitation

You can purchase digital transceivers specified for different distances, with 10, 40, and 70 km being common for gigabit Ethernet systems. These distances assume no splitting loss (gigabit Ethernet is a point-to-point system). With splitting loss, each two-way split costs nominally 3.5 dB of signal. At 1310 nm, the loss of the fiber is 0.35 dB/km, so each split costs 3.5/0.35, or 10 km of distance. Add 2 dB for connector loss, as we did in the analog transmission case. This costs 2/0.35, or nearly 6 km of distance.

Table 19.3 shows that splitting the signal reduces the distance you can cover very rapidly. Each two-way split costs 10 km in distance covered, and we are not allowing for worst-case losses in the splitters. Higher split ratios are desirable, because higher split ratios correspond to fewer fibers coming into the hub or headend. Each fiber that returns to the hub or headend must be managed and must be repaired if a cable is cut. Furthermore, each cable returning to the headend must be terminated, either at an EDFA for broadcast fibers or in a router for digital baseband systems.

Some PONs are limited to 10 km or so in order to allow for a reasonable amount of splitting. Splits of up to 64-way are common, though usually at speeds much lower than 1 Gb/s, the speed upon which Table 19.3 is based. As speed is reduced, distances and/or splits can increase. Reducing loss or increasing power by 3 dB will allow a doubling of speed or will permit an additional distance of about 8.5 km. Note from Table 19.2 that if the PON is carrying analog signals, the number of splits is constrained to 32 or fewer.

Table 19.3 Approximate distance limitation of common GBICs, with splitting (gigabit Ethernet)

Split	*Nominal splitting loss*	*Distance, km (1,310-nm, 10-km GBICs)*	*Distance, km (1,310-nm, 40-km GBICs)*	*Distance, km (1,550-nm, 70-km GBICs)*
4 way	7 dB	—	14	44
8 way	10.5 dB	—	4	34
16 way	14 dB	—	—	24
32 way	17.5 dB	—	—	14
64 way	21 dB	—	—	4

19.9.3 Standards for Fiber-Deep Systems

There is one complete standard for fiber-deep systems at the time of this writing and one activity that commenced in 2000, aimed at defining a second standard.

Full-Service Access Network (FSAN)

A standard has been developed for APONs, under the auspices of a group known as the *Full-Service Access Network* (FSAN). Formed in 1995 by seven telecommunications companies, it has grown to a membership of 21 carriers worldwide, including BellSouth, Bell Canada, SBC, Verizon, Qwest, NTT, SingTel, and Telstra. Its specifications have been accepted by the ITU as G.983.

FSAN provides for fiber to the home (FTTH) or business (FTTB) as well as to the curb (FTTC — typically serves fewer than 16 subscribers) and to the cabinet (FTTCab). FTTCab serves more subscribers from one central terminating point.

FSAN specifies a maximum distance from the central office to the fiber termination of 20 km. The maximum optical split is 32. Downstream data transmission is on 1490 nm at a speed of either 155.52 or 622.08 Mb/s. Upstream transmission is at 1310 nm at a speed of 155.52 Mb/s. There is no video provided for in the basic specification, G.983.1. An extension, G.983.3, adds video at a wavelength of 1550 nm. This is sometimes called a BPON (broadcast passive optical network).

802.3ah

The other layer 2 protocol commonly used in PONs is Ethernet. Systems based on Ethernet are known as EPONs. Ethernet is by far the most popular layer 2 protocol used in local area networks such as those found in most offices. Ethernet uses variable-length packets, with the packet size ranging from 64 to 1500 bytes.

As of this writing, there is an industry consortium, under the auspices of the IEEE 802.3 committee, that is defining a standard. The group is known as 802.3ah, or, informally, 802.3 EFM, for *Ethernet in the first mile*. Several companies developed Ethernet-based systems in advance of a standard.

19.10 Summary

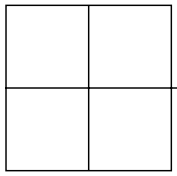
Besides using frequency division multiplexed analog transmission of multiple carriers, it is possible to use digital (binary, or on/off) modulation of lasers, modulating them with a time division multiplexed baseband digital signal. Both forms of optical modulation may be used on the same fiber if different wavelengths are multiplexed. One application of such digital systems is intercity and metro data transmission, but it is becoming more common also to use it, often in

conjunction with analog FDM modulation, to deliver signals via fiber directly to homes and businesses. This is known as fiber to the home (FTTH) and fiber to the business (FTTB). Some systems deliver fiber signals to the curb (FTTC). Collectively these are known as fiber-deep systems.

Fiber-deep systems can be differentiated by the layer 2 technology used for data transmission: ATM or Ethernet. They are also differentiated by whether or not they use active elements in the plant. Typically such systems deliver baseband data and modulated video signals to a box on the side of the home, which converts signals back to the more familiar electronic form. The most common data interface is Ethernet, whereas telephone interfaces are predominately analog, and broadcast television is delivered as in conventional cable plant.

Endnote

1. Institute of Electrical and Electronics Engineers, *Standard 802.3, 2000 Edition*, section 38.2.4.



Chapter 20
Network Reliability and Availability

20.1 Introduction

Chapter 17 described the network requirements of certain classifications of services that can be offered over broadband FDM distribution networks. Among those are transmission quality, effective bandwidth, *service reliability* (the probability that a system will survive without interruption for a defined period), *outage rate* (the average rate at which service interruptions occur), and *availability* (the percentage of time that service is available). Transmission quality was treated in detail in Chapters 10–16, and effective bandwidth was covered in Chapter 18. This chapter will deal with the calculation of network reliability and service availability. It will also discuss how those parameters vary as a function of the topology of the distribution system, among other factors. Finally, it will explore the difference between true availability and that experienced by users of a particular service.

Service interruptions can result from a variety of causes, including equipment failure, commercial power problems, interfering signals, or blocking due to inadequate circuit capacity, as discussed in Chapter 17. This chapter will deal with outages caused by either equipment or powering. Chapter 4 dealt with data traffic engineering, while Chapter 16 dealt with upstream interference issues.

20.2 History and Benchmarking

Logically, service availability should be just that — the net availability to the end user, irrespective of the cause of any interruptions. Historically, however, different industries have used different measures of their own service integrity.

20.2.1 Telephony

As we discussed in Chapter 17, the most widely quoted service availability requirement is 99.99% for local, wired, voice telephony service. It is important to review, however, how the exchange carriers define this parameter:

> The service availability objective for the subscriber loop is 99.99%, which corresponds to 0.01% unavailability or 53 min/yr maximum downtime. This objective, incorporating all network equipment between the local switch and the network interface (excluding the local switch and customer premises equipment), refers to the long-term average for service to a typical customer.[1]

This is not the same as end user service availability. It excludes, for instance, availability of dial tone from the switch, for whatever reason, or availability of interswitch or long-distance circuits. It also excludes any wiring within the home or terminal equipment problems. Though this may be understandable from a local network operator's point of view, the history of the cable television industry has been that in-home problems are the largest single cause of service outages.[2]

Second, this is a goal for performance that is averaged over both the entire customer base and over time. Thus, a significant minority of customers may have consistently unreliable service, yet the carrier meets the service goal if the majority of customers have few problems.

Third, carriers do not count "unavailability due to the loss of both primary and back-up powering"[3] in cases where there is powered equipment between the central office and home, such as fiber-in-the-loop (FITL) configuration. The importance of this exception depends, obviously, on the integrity of the field powering systems. It has been the experience of cable television network operators, however, that utility power interruptions cause more outages than all equipment failure classifications combined.

Finally, availability includes not only failure rates but also the time required to restore service. The problem is that, though a cut telephone drop wire (as an example) causes service to be unavailable, the network operator will typically not be aware of the outage until the customer tries to use the telephone and discovers the line is dead, then reports the outage to the phone company. The phone company calculates the length of the outage, for purposes of measuring compliance with the availability goal, from when it is *reported*, not from when it actually begins.[4] Arguably, therefore, the actual outage times may far exceed the reported times. As the traditional telephone network evolves by deploying fiber and performance-monitoring equipment closer to homes, this disparity will presumably decrease.

20.2.2 Cable Television

Although the cable television industry has not historically had formal goals covering availability or outage rates, CableLabs undertook a study of field reliability and subscriber tolerance of outages under the auspices of its Outage Reduction Task Force. Its report, published in 1992,[5] is discussed in detail in Chapter 17. In summary, the report set as a failure rate target that no customer

should experience more than 2 outages in any 3-month period, with a secondary standard of no more than 0.6 outage per month per customer.

As with the telephony goal, it is important to review what is included in the definition of an outage:

- Only multicustomer outages are included, thereby excluding all drop, NID, set top terminal, or in-house wiring problems.
- Single-channel outages are included, thereby including headend processing equipment.
- Only signal outages, not degraded reception, are included.
- All elements of signal distribution between headend and tap are included, including the effects of commercial power availability.
- Although the language refers only to outages that are "experienced by customers," the software distributed with the report calculates all outages, and therefore, it can be assumed that customer awareness of outages is not important to the definition, even though the goal was based on customer experience.
- The "no customer shall experience" clause in the goal clearly indicates that a worst-case, not average, analysis is called for.

Based on an estimated mean time to restore service of 4 hours, CableLabs translated 0.6 outage per month into a minimum acceptable availability of 99.7%.

The equivalence between customer-perceived individual outages and distribution system outages is tenuous at best. On the one hand, some outages may be detected and repaired without some subscribers being aware of the outage; on the other hand, individual subscribers may have in-home problems that will not be included in the preceding definition. Nevertheless, CableLabs chose 0.6 multicustomer outage per month as its benchmark for distribution system outage modeling.[6]

In the field measurements portion of its study, the task force gathered actual field failure analysis reports from some of its members and used those to predict failure rates of certain classifications of components (discussed further later) and to develop models that could be used to predict the failure rates of arbitrary network architectures. The accuracy of the raw data used in this type of analysis is limited by the analytic skills of the field technician and by its completeness. For instance, it is not obvious that a technician would think to report a brief outage caused by a routine maintenance procedure or by a brief power outage that did not cause a formal trouble call to be entered into the dispatch system. Second, the gathered data is necessarily representative of typical cable systems of that time and did not include modern HFC systems. Finally, cable systems (at least in 1992) routinely reported only *distribution system outages*, not *individual customer outages*, in their reports.

Table 20.1 Comparison of Bellcore and CableLabs outage definitions

	Factors Included in Definition of Outage						
Outage Definition	*Centralized Processing/ Switching*	*Distribution Network*	*Distribution Power*	*Drop*	*Network Terminal Equipment*	*In-Home Wiring*	*Analysis*
Bellcore		X		X	X		Average customer
CableLabs	X	X	X				Worst-case customer
Customers' Experience	X	X	X	X	X	X	

20.2.3 Cable Television Versus Telephony Reliability and Availability Definitions

Clearly, the CableLabs and Bellcore definitions of reliability and availability are not equivalent, and the numbers should be not compared directly. Table 20.1 illustrates the approximate inclusion of each industry's outage definition compared with those elements that actually affect customer service.

As a final note, *availability* is traditionally the most important parameter applied to telephone services though CableLabs found that the *rate at which outages occurred* was most important to television viewers, with the length of the outage (and thus availability) of only secondary importance.[7]

20.3 Definitions and Basic Calculations

It is important to have a clear understanding of the terminology used in calculating reliability-related parameters. Thus, the following definitions will be used throughout this chapter.

- *Component failure rate*. The failure rate of a component or system is defined as the statistical probability of failure of any one of a similar group of components in a given time interval. The measured period should be contained within the normal service life of the device — that is, failure rate is not intended to measure time to wear-out, but rather "midlife" random failures. For the purposes of this chapter, we will generally express failure rates on a yearly basis.

  Failure rates of individual electronic components (capacitors, resistors, semiconductors, and so on) are generally determined by their manufacturers

using carefully designed tests that are conducted on a sample of products selected at random from the production line.

Failure rates of system-level components such as amplifiers, taps, connectors, and set top terminals can be determined theoretically based on the failure rates of their internal electronic parts and the way in which they are interconnected, or they can be determined based on actual failure data from field-installed units. We have a strong preference for the latter since the installed failure rate is influenced not only by design but also by craft issues, environmental conditions, and how the unit is powered.

When field observations are used to determine failure rate, the appropriate formula is

$$\lambda = \frac{k}{nt} \tag{20.1}$$

where

λ = the annual failure rate
k = the number of observed failures
n = the number of items in the test sample
t = the observation time in years

The failure rate can be converted to other units as required. For instance, the failure rate per hour is the annual failure rate divided by 8,760.

- *System failure rate—simple nonredundant connection*. Where n components are connected in such a way that the failure of any one of them will lead to a failure of some defined system or function (such as a subscriber service outage or a single-channel failure), and the failures do not have a common cause, the resultant failure rate of the system is given by

$$\lambda_{\text{system}} = \sum_{i=1}^{n} \lambda_i \tag{20.2}$$

where

λ_{system} = the net system failure rate
λ_i = the failure rate of component i
n = the number of components, any of whose failure will cause a system failure

Note that this formula is not exact since it assumes that none of the failures overlap in time. Thus, it is accurate only when the failure rates and repair times are such that multiple, independent, simultaneously occurring outages

are statistically insignificant. Generally, this is a good assumption for well-designed and managed cable systems.

- *System reliability — simple nonredundant connection*. System reliability is the probability that the system will not fail during some specific period. In general, reliability can be calculated once the failure rate is known using

$$R(t) = e^{-\lambda t} \tag{20.3}$$

where λ and t are in consistent units (that is, failure rate per year and years or failure rate per hour and hours).

The result of this calculation may be less than intuitive. For instance, if the annual failure rate of a given system were 1 (meaning that if we measured a large number of similar systems, we would detect a number of failures per year equal to the number of systems), the probability of any single system surviving one year without a failure is $e^{-1} = 0.37$.

Where components or subsystems of known reliability are connected in a configuration where the failure of any component leads to a system failure, the net system reliability can be calculated using

$$R_s(t) = R_1(t)R_2(t)\ldots R_n(t) \tag{20.4}$$

where

$R_s(t)$ = the net system reliability over period t

$R_1, R_2, \ldots, R_n$ = the reliabilities of the individual components, measured over the same period

- *Mean time between failure (MTBF)*. The predicted time between failures is simply the inverse of the failure rate, or

$$MTBF = \frac{1}{\lambda} \tag{20.5}$$

where *MTBF* and λ are in consistent units.

Generally, this chapter will express MTBF values in hours. When manufacturers provide predicted MTBF data, it is also generally given in those units. As with failure rate, MTBF is not a measure of useful service life. In fact, many components have an MTBF that exceeds their wear-out time.

- *Mean time to restore (MTTR)*. When a failure does occur, the mean time between failure and restoration of the defined function is the MTTR. For the purposes of this chapter, MTTR values will usually be expressed in hours.

Note that MTTR is sometimes used to express mean time to *repair* rather than mean time to *restore*. We use the latter definition here to distinguish between when service is restored and when the failed component is actually

repaired. Since it is more common in cable television to restore service after a failure by swapping a good for a failed network component, this is more accurate when our goal is to determine service reliability.

- *Availability (A)*. The ratio of time that a service is available for use to total time is known as availability. It can be calculated from MTBF and MTTR (expressed in consistent units) using

$$A = \frac{MTBF}{MTBF + MTTR} \tag{20.6}$$

- *Unavailability (U)*. Unavailability is the ratio of time that a service is unavailable to total time and is, by definition, equal to $1 - A$.
- *Outage time* (T_U). Outage time is the amount of time that the network is unavailable during a defined period. In general, the outage time during some period t is equal to the unavailability multiplied by t. It is common to express outage time in minutes per year, which can be calculated as follows:

$$T_U \text{ (min /yr)} = (\text{minutes/year})(1 - A) = 525,600(1 - A) \tag{20.7}$$

20.4 Effects of Redundant Network Connections

The preceding simple calculations apply to situations in which the failure of any single component leads to the failure of the defined system or function. In a redundant connection, either of two (or more) components can provide a function should the other(s) fail. Thus, the system user experiences a failure only if the redundantly connected components simultaneously fail. As discussed in the previous chapter, a common example in fiber architectures is a ring configuration where signals can travel by either of two cable routes between headend and hubs.

- *Unavailability*. When two network sections are connected in a redundant configuration, the equivalent unavailability is the product of the unavailabilities of the two individual sections. Since, for each section, the unavailability is simply the number of outage minutes per year divided by the total minutes in a year, the equivalent unavailability is

$$U_R = \frac{(T_{U1})(T_{U2})}{(525,600)^2} \tag{20.8}$$

where

U_R = the net unavailability

T_{U1} and T_{U2} = the yearly outage times (in minutes) of the redundantly connected sections

Converting this back into equivalent yearly outage minutes for the combined section:

$$T_{UR} = 525,600U_R = \frac{(T_{U1})(T_{U2})}{525,600} \tag{20.9}$$

- *Failure rate*. The equivalent failure rate is the sum of the failure rate of the first feeding segment multiplied by the probability of the second section failing before the first can be restored, plus the failure rate of the second section multiplied by the probability of the first section failing during an outage in the second section.

  The probability of section one failing during an outage in the redundantly connected section two is

$$P_{F1} = 1 - e^{\frac{-(\lambda_1)(MTTR_2)}{8,760}} \tag{20.10}$$

  where

  P_{F1} = the probability of failure of section one
  λ_1 = the failure rate of section one per year
  $MTTR_2$ = the period (in hours) over which the probability of failure is to be measured, namely, the MTTR for a failure in network section two

  Thus, the net failure rate when two sections join in a redundant feed is

$$\lambda_R = \lambda_1\left[1 - e^{\frac{-(\lambda_2)(MTTR_1)}{8,760}}\right] + \lambda_2\left[1 - e^{\frac{-(\lambda_1)(MTTR_2)}{8,760}}\right] \tag{20.11}$$

20.5 Absolute Versus User-Perceived Parameters

The preceding equations yield reliability-related performance parameters for the network, independent of usage. For applications that require continuous use of the network, the calculated failure rates and availability of the network are the same as those experienced by a user. For most services, however, usage of the network by individuals is only partial. For instance, most families watch television about 4 hours per day, whereas telephone users may access the public switched telephone network for only a half hour daily.

The importance of partial usage is that some outages will occur and service be restored before all potential users are affected. Thus, the average network reliability, as experienced by occasional users, will be higher than for full-time users. However, the relationship between the network outage rate and avail-

ability and that experienced by occasional users is complex and affected by not only average aggregate usage time but also by the pattern of usage.

As an example, if a television subscriber watches 4 straight hours of television every evening, but at no other time, and the network MTTR is 1 hour, the probability is that he or she will experience slightly more than one-sixth of all outages. On the other hand, a person who uses the same network to make a 1-minute phone call every 6 minutes (say, a 24-hour business using a telephone modem to do credit card verifications) will experience every outage (and, in fact, will perceive a greater outage *rate*, since the restoration time for each outage will affect 10 attempted phone calls) even though his or her total daily usage time is the same.

A further complication is in the user's response to a failure and how we quantify the modified usage pattern. For example, assume that a user normally averages one phone call every 4 hours. If the user attempts to make a call and finds the phone not working, he or she may try again every few minutes until successful. If the outage lasts an hour, the user will have experienced many more failures than during his or her normal usage pattern. At the other extreme, a television viewer who would normally watch an evening's programming may, upon discovering an outage, abandon viewing for the entire evening and thus not be aware that the outage lasted only a few minutes. Again, the viewer's perception of the network's availability is affected by his or her changed usage pattern. A legitimate question for reliability modeling is whether these failure-induced retries are counted as independent failures.

For a hypothetical fiber-to-the-curb, fully status-monitored network, one researcher has calculated that a telephone user making 10-minute calls spaced 100 minutes apart (roughly a 9% utilization factor of the network) would experience 30.2 of the network's 53 minutes of unavailability every year.[8] If any portion of the network is unmonitored, the outage time experienced by the customer will increase and could, in fact, exceed the Bellcore-defined outage time for the network!

For cases where usage patterns are not modified as a direct result of outages, we can estimate perceived outage rates and service availability as follows.[9]

First, define usage pattern for a service consisting of an average *call length* as T_C, and an average *call cycle* time (from the start of one call to the start of the next call) as T_S. To keep things simple, all times will be expressed in hours.

The probability of a service failure during a call period T_S can be calculated from Equation (20.3) (the probability of failure is 1 minus the reliability):

$$P_f = 1 - e^{-\lambda T_s} \tag{20.12}$$

where T_S is in hours and

λ = the service failure rate per hour

If we multiply the probability of failure per call cycle by a scaling factor that takes into account both call length and service restoration time, we can calculate the probability of failure per attempted usage of the network:

$$P_C = (1 - e^{-\lambda T_s})\left(\frac{T_C + \text{MTTR}}{T_S}\right) \tag{20.13}$$

where T_C and MTTR are both in hours.

The total number of attempts to use the service per year can be calculated using

$$B = \frac{8,760}{T_S} \tag{20.14}$$

from which we can calculate the user-experienced yearly failure rate:

$$\lambda_{\text{EXP}} = P_C B \tag{20.15}$$

The average length of experienced outages is

$$T_O = \frac{(T_C)(\text{MTTR})}{T_C + \text{MTTR}} \tag{20.16}$$

so that the total experienced outage time per year is

$$T_{\text{EXP}} = (P_C)(B)(T_O) \tag{20.17}$$

Finally, the experienced availability (that is, 1 minus the ratio of outage time to attempted usage time) can be calculated using

$$A_{\text{EXP}} = 1 - \left[\frac{(P_C)(T_O)}{T_C}\right] \tag{20.18}$$

As an example, assume that a network fails once every 3 months (MTBF = 2,190hours, $\lambda = 0.000457$ per hour) and that the MTTR is 2 hours. Based on this data, an application that used the network continuously would experience an outage rate of 4 per year, a yearly outage time of 8 hours, and an availability of 0.999088.

A television viewer using the network 4 hours per day ($T_C = 4$, $T_S = 24$) would experience a yearly outage rate of 0.9953, a yearly outage time of 1.327 hours, and an effective availability of 0.999091. Thus, although using the net-

work only one-sixth of the time, the viewer experiences roughly one-fourth the outage rate but nearly the same effective availability as a 24-hour user.

A telephony user who makes a 10-minute call every 4 hours ($T_C = 0.16667$, $T_S = 4$) would experience a yearly outage rate of 2.1665, a yearly outage time of 0.333 hour, and an effective availability of 0.999087. Thus, even though the telephone user makes use of the network for a lower percentage of the time, he or she experiences about the same effective availability but a higher failure rate than the television viewer because of his or her more frequent attempts at access.

As can be seen, network effective availability — the probability that service is available — is quite insensitive to usage patterns, whereas experienced outage rate is closely tied to both call length and call rate.

20.6 Network Analysis

Once the basics of reliability are understood, they can be applied to the analysis of actual and proposed networks. The steps to analysis include breaking the network into manageable sections, properly assigning the components to the sections, determining the failure rates of individual network components, estimating the service restoration times for failures at various network levels, and calculating the various reliability parameters. Though the principles of the process will be described here, in practice, the calculations are much more easily performed using a computer, which also allows the engineer to assess the effects of changes in topology, operating practices, and component reliability.

20.6.1 Logical Subdivision of the Network for Analysis

The first step is subdividing the network into segments. The degree to which this is required depends on the level of analysis required. For instance, if the only parameter of interest is the failure rate and availability as experienced by the most distant subscriber, then the many branches in the coaxial distribution system can be ignored and only the longest cascade analyzed in a single calculation.[10] As a minimum, however, any network branches that form redundant structures must be separated, analyzed separately, and then used as variables in the equations in Section 20.3 to determine the effective performance of the redundant structure.

For those desiring a more thorough analysis, a separate analysis of each physically separate branch and calculation of reliability data at every customer port will allow additional data such as total failures and the performance as experienced by the average as well as by the most disadvantaged customers. It also allows analysis of the contributions of various network sections to overall failure rate and unavailability. The analysis given at the end of this chapter was done using a large spreadsheet that allowed a very detailed analysis.

In order to limit the amount of data to be analyzed, it is common to analyze only a typical subsection of a large network. In the case of an all-coaxial network, this might be a single trunk, whereas in the case of an HFC network, it is likely to consist of a single fiber-fed node, along with the fiber trunking structure feeding that node.

Figure 20.1 shows a typical example. In this case, a sheath ring fiber structure is used between the headend and the node receiver, while coax cascades of up to five amplifiers are used beyond the node. Should only the performance to the last user be required, then analysis of all but the longest cascade can be omitted. Only the downstream optical structure is shown in the Figure. See Section 20.6.6 for a discussion of the calculation of two-way structures.

Figure 20.2 shows the logical division of this network into segments for analysis. Note that the number assigned to each section identifies a unique structure, whereas the letter denotes the instance of that structure. Thus, all four of the segments numbered five are similar, and only a single analysis of that structure is required. Since a single headend optical transmitter feeds both downstream fibers between the headend and node, it is part of segment 1A. Separate optical receivers are used in the node, however, so each of the segments numbered 2 includes both the fiber and connected node receiver.

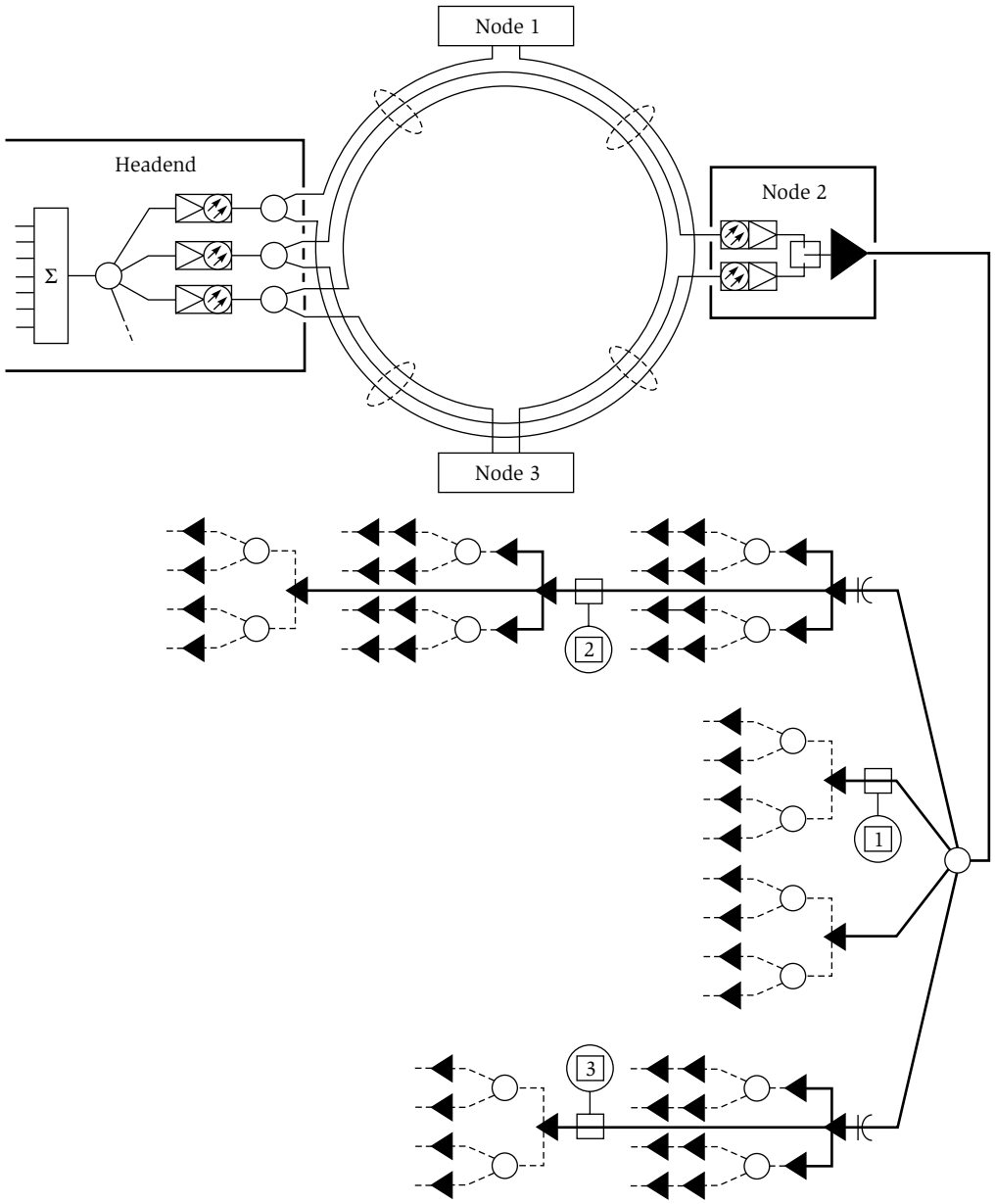
20.6.2 Assigning Network Components to Subdivisions

Sometimes, network components' effect on transmission integrity may be determined by other than their physical placement in the network. Power supply 1 in Figure 20.2 is physically located in section 5A, but powers the node and the first amplifier in both sections 5A and 5B. Therefore, for purposes of calculating reliability parameters, it should be included in section 3A. Similarly, power supply 2 is physically located in section 6A, but must be included in section 4A since it powers the amplifier that feeds sections 7A and 6A, and its failure would affect the signals in all the downstream branches.

With the components properly assigned to network segments, based on their effect on operation, the components in each section should then be listed in a column as a first step to segment analysis.

20.6.3 Determining Component Failure Rates

In order to estimate the performance of the defined sections, the field failure rate of each network component type must be determined. Manufacturers' design analysis, based on the MTBF of the individual components used and their temperature/voltage stress levels, often have limited relationship to failures of field-installed equipment where such factors as voltage transients and



Note: Only downstream F/O network shown for simplicity. Taps omitted for clarity, untapped cable shown solid, and tapped cable shown dashed.

Figure 20.1 Topology of typical HFC node and sheath ring trunk.

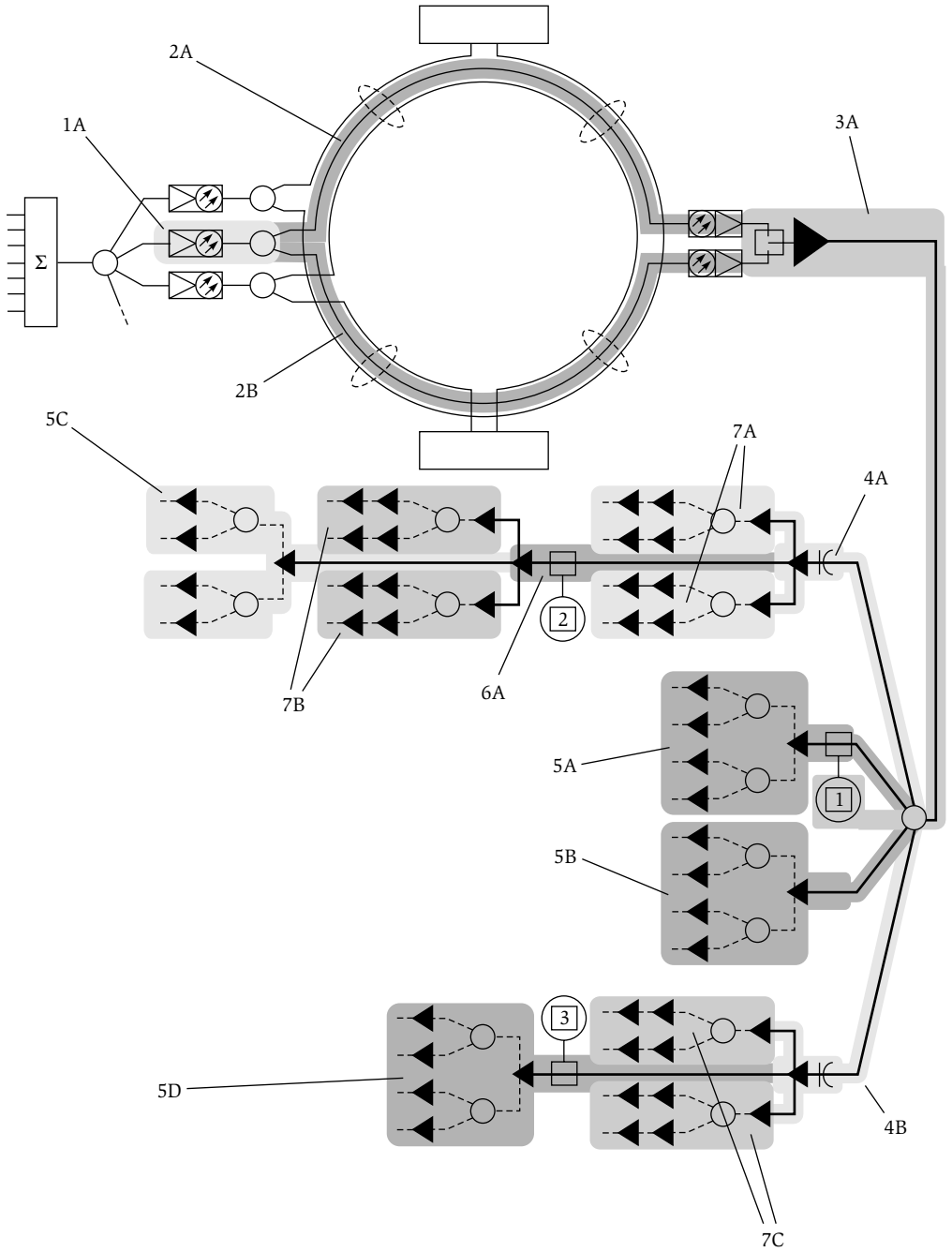


Figure 20.2 Division of network into segments for analysis.

the skill levels of field personnel may dominate inherent failure rates. Also, a failure of a series-connected component may or may not cause a loss of service to all downstream subscribers. For instance, a tap port may fail, causing a loss of service to one customer but not disrupting the through-signal to subsequent taps.

With those caveats, it is instructive to examine failure rates of a few key components.

- *Commercial power*. CableLabs' task force recorded average commercial power failure rates of 30% (all failure rates are on a yearly basis unless otherwise noted), meaning that the probability of a commercial power interruption at any given power supply is 0.3 each year. Since this data is based on outages caused by power failures, it is to be presumed that this Figure relates to *unprotected* commercial power outages rather than all power outages. Also, given that their data was derived from reported system outages, it can be presumed that many short interruptions were not included in the data.

  By contrast, the Network Reliability report quotes from a study showing average commercial power *unavailability* of 370.2 minutes per year.[11] They did not report on the rate at which outages occurred, however. The net unavailability with varying durations of standby power capacity is shown in Table 20.2.

  From this data, it is readily apparent that even standby battery capacity of several times the 2 hours normally designed into cable television standby power supplies will result in unprotected outages well in excess of Bellcore's 53-minute availability guideline. Thus, normal standby powering will need to be supplemented by other means (such as built-in generators or status monitoring combined with deployable mobile generators) for high-availability services.

  It is the experience of one of the authors, however, that the reliability of commercial power varies widely from location to location so that Table 20.2 can be taken only as a measure of average conditions. Accurate network reliability predictions must depend on locally viable commercial power data unless networks are so hardened against commercial power outages as to make the issue irrelevant.
- *Power supplies*. Independent of the availability of commercial power is the reliability of the power supplies themselves. CableLabs' data has suggested a failure rate of only 3%. Though such a rate may be achievable by a nonstandby

Table 20.2 Average commercial power unavailability in the United States

Average Service Unavailability (minutes/year)	*Zero Backup*	*4-Hour Backup*	*8-Hour Backup*	*12-Hour Backup*
	370.2	192.4	153	128.5

power supply, it is the industry's common experience that standby power supplies are often much less reliable (though the overall reliability of commercial power plus standby power supply will exceed that of a nonstandby supply by itself). Standby supply reliability will be dramatically affected by the age and condition of the batteries, as well as by the level of routine maintenance in the system.

- *Amplifiers*. CableLabs found field failures occurring in trunk amplifiers at the rate of 10%, whereas failures in the simpler line extenders occurred at a 2% annualized rate. "Distribution" amplifiers used in modern HFC networks have a complexity somewhere between these extremes, and 5% is often used in modeling HFC networks though upgraded designs may justify lower values. Werner of TCI has quoted "empirically derived" failure rates for distribution amplifiers as low as 1.75%.[12] Merk has suggested that operators use failure rates of 0.5% for trunk amplifiers and 0.15% for line extenders, based solely on warranty repair data.[13] This range of values illustrates how difficult it is to be confident of failure rates and how competent engineers can arrive at rates that vary by more than 10:1.
- *Optical transmitters*. Werner and Merk are consistent in reporting 2.3% downstream transmitter failure rates. If CableLabs' failure rates for other headend equipment of 20–30% are to be believed, then these values are suspiciously low. On the other hand, the base of equipment that went into the latter study varied in age, whereas optical transmitters are likely to be much newer given the relatively recent widespread rollout of HFC systems.

  Werner rates upstream transmitters identically to downstream, whereas Merk suggests that upstream transmitters experience a failure rate of only 0.9%. Given that most upstream transmitters operate at much lower powers, do not require a thermoelectric cooler, and do not require a high-power driver, there is some logic for believing a lower upstream value.
- *Optical receivers*. When evaluating optical receivers, it is important to distinguish between the receiver itself and the entire optical node. Since nodes often include full trunk amplifier stations, it is advisable to rate the receiver independently of its host station unless it is a stand-alone unit. Werner's value here is 1.7%, whereas Merk uses 0.7%, with both authors rating upstream and downstream receivers essentially identical.
- *Passive devices (taps, power inserters, splitters, directional couplers)*. MTBF data supplied to CableLabs by manufacturers suggested that passive device failure rates should be on the order of 1%. CableLabs' field data suggested that actual numbers were closer to 0.1% — a rare case where actual experience was superior to predicted. Werner's data is even more optimistic at 0.04%. Merk suggests the use of values between 0.07% and 0.26%, depending on the device.
- *Coaxial connectors*. Neither CableLabs nor Merk suggests failure rates for coaxial connectors though CableLabs' "cable span" failure rate may be a

composite of the actual cable and the attached connectors. Werner suggests a failure rate of 0.01%. Although this may seem insignificant, it must be remembered that the signal must pass through at least two connectors for every series-connected device in the system so that a cascade of 5 amplifiers, 1 power inserter, 2 splitters, and 25 taps will include 66 connectors.

The preceding values are for the connectors used for solid-sheathed aluminum cable. From experience, the failure rate (as measured by outages and associated trouble calls) for drop connectors is much higher. Based on actual experience operating numerous systems, it is suggested that 0.25% be used pending better industry statistics. As with distribution connectors, a typical drop system will include several series-connected F connectors. Among all the components considered, the greatest actual field variation may be in the performance of drop connectors owing to both craft sensitivity and the wide range of available connector quality choices.

- *Fiber-optic cable*. The modeling program provided with CableLabs' report suggests the use of 3% failure rates for fiber cable spans (regardless of length). Merk suggests the use of Bellcore data of 0.44% per mile, whereas Werner suggests the use of 0.44% (presumably per span). Hamilton-Piercy suggests that Bellcore's data on fiber failures is equivalent to 0.3% per span per year but uses his company's experiential value of 1.5% per span.[14] Clearly, these are widely varying numbers that reflect several variables.

  For one thing, the Bellcore data was based on analysis of "FCC-reportable" outages only, meaning those that affect at least 30,000 customers.[15] Thus, most instances of fiber damage on the subscriber side of switches went unreported, whereas the failures occurred in telephone interoffice cables that are placed in well-protected trenches because of the high volumes of traffic being carried. Fiber serving small pockets of customers is more economically installed and more likely to be damaged as a result. For example, most of the damaged cables reported by telephone companies were located 4–5 feet underground, well below the depths used in most cable television construction.

  Second, the lengths of the spans considered vary from 5 miles or less for typical suburban HFC architectures to 50 miles or more for long-distance telephone links. Finally, available failure data may have been difficult to coordinate to the total length of the failed cable.

  Given that most fiber cable failures are due to dig-ups and should increase proportionally with distance, a value of 0.1% per mile is suggested pending better data. This is roughly equivalent to Hamilton-Piercy's value if 15-mile average link distances are assumed, or Werner's value if 4.4-mile spans are assumed.

- *Coaxial cable*. Coaxial distribution cable might be thought to suffer similar failure rates. CableLabs' field data, however, suggests 3% annual failures. Even if it is assumed that spans are a half-mile long (about the longest expected between trunk amplifiers), this would be the equivalent of 6% per

year per mile, or 40 times the optical cable failure rate. Werner's 1996 data suggests the use of 0.23% — less than one-tenth the failure rate. Based on the fact that fiber cable is generally armored and therefore both stronger and not as subject to rodent damage, it seems reasonable to use a value of 0.2% per mile for distribution coaxial cable or double the risk factor for fiber cable.

Drop cable is not as well protected, nor is it made to the same quality standards as distribution cable. Analysis of trouble call rates and causes suggests that annual failure rates of drop cables may be in the 1–2% region.

- *Set top terminals and other operator-supplied terminal equipment*. The failure rates of terminal devices will obviously vary considerably with complexity and age. New, nonaddressable terminals will have very low failure rates, whereas first-generation addressable products often failed at several percent per month. Few authors who have modeled cable television failure rates have included terminal data since CableLabs' definition excludes individual subscriber outages. Based on field failure data from several systems over a number of years, we suggest that a default value of 7% be used, lacking more device-specific data. For purposes of analysis, this failure rate is taken to include nonhardware failures such as the failure of an addressable terminal to receive an enabling command, data entry errors that result in unintended de-authorizations, and power failures that cause the terminal to lose its authorization status for some time after power restoration. From the subscribers' viewpoint, these are still service outages.
- *Active network interface devices (NIDs)*. Where services terminate at the outside wall of residences, the equipment is subjected to greater weather exposure but may be rigidly mounted and is not subject to handling damage. Bellcore suggests that NIDs containing active termination equipment, whether for video or telephony service, should have a net unavailability of about 26 minutes per year.[16] If it is assumed that individual customer outages are restored in a mean time of 8 hours, this is equivalent to a 5.4% yearly failure rate.
- *Headend/rack-mounted equipment*. CableLabs' initial data indicated very high (20–30%) failure rates among headend equipment. In general, however, a failure rate of 5%, averaged over all classifications of headend equipment, is more in line with operators' experience and will be used in our example calculations.

  In the case of telephony, headend units are required to interface between the ports on the switch and the HFC distribution system. Although the term often has a more precise meaning in different contexts, host digital terminal (HDT) will be used to refer to all the headend equipment between the switch and broadband distribution system, considered as a subsystem. Although Bellcore has not predicted the failure rate of HDTs, they have assigned 10 minutes of annual unavailability to this subsystem. On the basis of an assumed 0.5-hour MTTR, the HDT is assigned an annual failure rate of 33%.

Table 20.3 Typical individual network component annual failure rates

Component	*Failure Rate*	*Component*	*Failure Rate*
Commercial Power (unprotected)	300%	Optical Receivers	1%
Power Supplies (nonstandby)	3%	Passive Devices	0.1%
Power Supplies (standby)	6%	Coaxial Connectors (distribution)	0.01%
Amplifiers (trunk or distribution)	4%	Coaxial Connectors (drop)	0.25%
Amplifiers (bridger modules or line extenders)	3%	Fiber-Optic Cable	0.1% per mile*
Optical Transmitters (downstream)	3%	Coaxial Cable (distribution)	0.2% per mile
Optical Transmitters (upstream)	1.5%	Drop Cable (length independent)	1.5%
Redundancy Switches	0.5%	Set Top Terminals/Modems	7%
NIDs	5.4%	Headend Equipment	5%
HDTs	33%		

* Since the major cause of failure is dig-ups and cuts, the failure rate applies to the entire cable and all the fibers within it. Therefore, no redundancy gain is realized by using separate fibers to deliver signals to a node, for instance, if they share a common cable sheath.

In summary, there is reasonable industry agreement on some average failure rates but widely diverging data on others. As the cable industry gains more experience with reliability as an essential element of network design, better data will likely become available for predicting performance. We have a bias toward carefully collected field data derived from systems operating in conditions similar to the system to be analyzed (regional differences in such factors as lightning strikes and the availability of laws requiring utility locates before commencing underground construction can have a major effect on outages). For the purposes of the example discussed later in this chapter, the annual failure rates shown in Table 20.3 will be used.

20.6.4 Estimating MTTR

Although the elapsed time between failures is a function of network architecture and the quality of the components used, the time required to restore service after

an outage occurs is almost entirely within the operator's control and is a function of staffing levels and training in addition to performance-monitoring systems.

Total time to restore service consists of four time segments:[17] (1) the time between the start of the outage and when the operator is aware of it, (2) the time for crews to respond to the outage location, (3) the time for the technician to identify the cause of the outage, and (4) the time to repair or replace the failed item and restore service.

Werner quotes average MTTR values varying from 66 to 243 minutes, depending on the diligence of the system operator. Though CableLabs did not put forth an industry goal, they did quote from a Warner Cable guideline calling for 2-hour maximum MTTR for damaged fiber-optic cables.[18] Data provided by the telephone industry to the FCC on fiber repairs showed that, in most cases, service was restored within a 2–4- or 4–6-hour window.[19]

Historically, many cable systems used manual trouble call routing, staffed during business hours only (except for major outages, storms, and so on), and operated many headends without any regularly present technicians. As larger, regional systems offering a wide variety of services have evolved, 24-hour, 7-day staffing of headends is no longer uncommon, and field service staffs often work 2 shifts, with on-call technicians for late-night situations.

As an economic choice, the response to outages is proportional to the number of customers affected. Large outages inevitably clog the customer service telephone lines and may lead to bad publicity since customers already affected by loss of service are doubly inconvenienced by being unable to report the situation.

Based on experience from a number of systems, the MTTR values in Table 20.4 will be used in the example discussed later in the chapter. To the extent that operators do better or worse than these values (which may be overly pessimistic for a well-organized system), the availability will vary from that calculated.[20] Factors such as traffic conditions, dispersal of personnel, training, availability of repair materials, and communications systems and utilization will have a major impact on MTTR.

Table 20.4 Typical cable system MTTR values

Failure Level	*MTTR*
Headend (24-hour, 7-day staffing)	0.5 hour
Fiber trunk cables	2 hours
Optical node components	1 hour
Distribution system coaxial components	4 hours
Individual customer drop or components	8 hours with remote monitoring 12 hours without remote monitoring (customer call-in delay)

20.6.5 Special Considerations Within Headends and Hubs

When Bellcore set up the availability standards for telephony, they specifically excluded the switch and everything on the network (nonsubscriber) side of the switch. CableLabs, on the other hand, specifically included headends to the extent that a single-channel outage is counted the same as a total system outage.

Those wishing to model cable system reliability factors may wish to consider some interim approach that accounts for all critical common equipment (such as combining amplifiers and headend power) but only some percentage of individual channel-processing equipment. Without such an approach, it can readily be shown that the failure rate of a modern 750-MHz fully loaded headend will be theoretically very high, yet the average viewer will be unaware of those failures if they happen to occur in channels he or she is not watching.

In previously published studies, one of the authors has proposed that viewers, in a typical viewing session, are exposed significantly to 10 channels (that is, they access and view programming on 10 channels in an evening of typical viewing), and he therefore modeled headend reliability based on the sum of the critical common equipment, plus all the equipment required to create 10 channels (satellite antennas, receivers, modulators, and so forth).[21] In the example discussed later, it has been assumed that the viewer is exposed to the potential failure of the approximately 30 channel-specific pieces of equipment needed to process the 10 channels viewed, plus 3 common combining amplifiers. For other services, different calculations of headend equipment reliability might be appropriate. As with equipment failure rates, there is little statistically valid field information on the number of channels viewed by typical subscribers.

A complete analysis of video headend-caused subscriber outages should include not only signal-processing equipment but control equipment, for instance, the telemetry generators that send enabling signals to terminals, addressability control computers, and (for two-way terminals) upstream data receivers.

20.6.6 Special Considerations for Two-Way Systems

Most of the distribution system is shared between upstream and downstream signals, including amplifier stations (though a detailed analysis might assign different failure rates to upstream and downstream modules within the station), power supplies, passive devices, and the cable itself. The optical portion of the plant, however, almost always uses separate fibers for upstream and downstream communications, and always uses separate upstream and downstream optical transmitters and receivers.

When modeling the optical trunking segments, therefore, both upstream and downstream transmitters and receivers are critical to two-way applications and must be included. Assuming the fibers share a common cable, however, the reliability of a pair of fibers will not differ significantly from a single fiber, and so the number would be the same as for one-way transmission. When calculating

the failure rate of the headend for a two-way application, it is also necessary to include equipment that processes both incoming and outgoing communications.

20.6.7 Calculation of Failure Rates and Availability

Once the structure of the network is defined and the field failure rates and repair times of the equipment used are known, the next step is to calculate the total failure rate of each segment. This is simply the sum of the failure rates of each of the elements in the segment (Equation (20.2)). Next, using Equations (20.6) and (20.7), it is possible to calculate the availability of each section and the total yearly minutes of outage.

If the only result desired is total system failure rate and availability to the most affected subscriber, the failure rates and outage minutes of each of the critical series-connected sections can be added. The net availability can then be calculated using Equation (20.7). Additional data can be derived from the analysis, if desired, by, for instance, calculating the number of subscribers affected by each outage, calculating the outage rate and availability to each subscriber, and/or calculating the cost of repair for each class of failure. Although this level of analysis is much more complex, it yields a significant amount of additional information about the performance of the network, including

- Reliability and availability to the average, as well as most affected, customer
- The total number of outages and the total number of customer outages (the summation of each outage multiplied by the number of customers affected)
- The total cost of outage repair
- The distribution of outages by size

This additional information is very helpful in making design trade-offs that include not only the cost of construction but also the required customer service staffing levels and predicted repair costs.[22]

20.7 Analysis of a Typical HFC Network

There have been many published analyses of the reliability and availability of specific network configurations. One of the most comprehensive compared 13 different architectures and found availability values ranging from 99.7% to 99.99% based on 0.3% fiber link failure rates and 10% amplifier failure rates (failure rates for other components were not given). Using 1.5% fiber link failures and 5% amplifier failures, two architectures were found to exceed 99.98% availability: a ring/ring/star and a double star.[23]

As an illustration of attainable reliability factors, as well as the degree of information available from a comprehensive analysis of failure rate data, we give a complete analysis of a typical HFC cable system.

20.7.1 Architecture

The analyzed system is typical of an early 1990s upgrade. The downstream bandwidth extends to 550 MHz, whereas the return bandwidth is limited to 30 MHz. Although it is logically a simple single star network, several nodes are served from each large fiber-optic cable leaving the headend, and the analysis accounts for this shared risk. Figure 20.3 is a simplified diagram of a typical node. As with Figure 20.1, the tap configuration is not shown though each of the dashed lines contains taps (and sometimes splitters and/or directional couplers). Even though not shown in the Figure, the taps and branching are included in the analysis. The distribution system extending from each node passes approximately 2,000 homes, with a basic penetration rate of 70%. The total number of homes served from the headend is 150,000, split among 75 similar nodes.

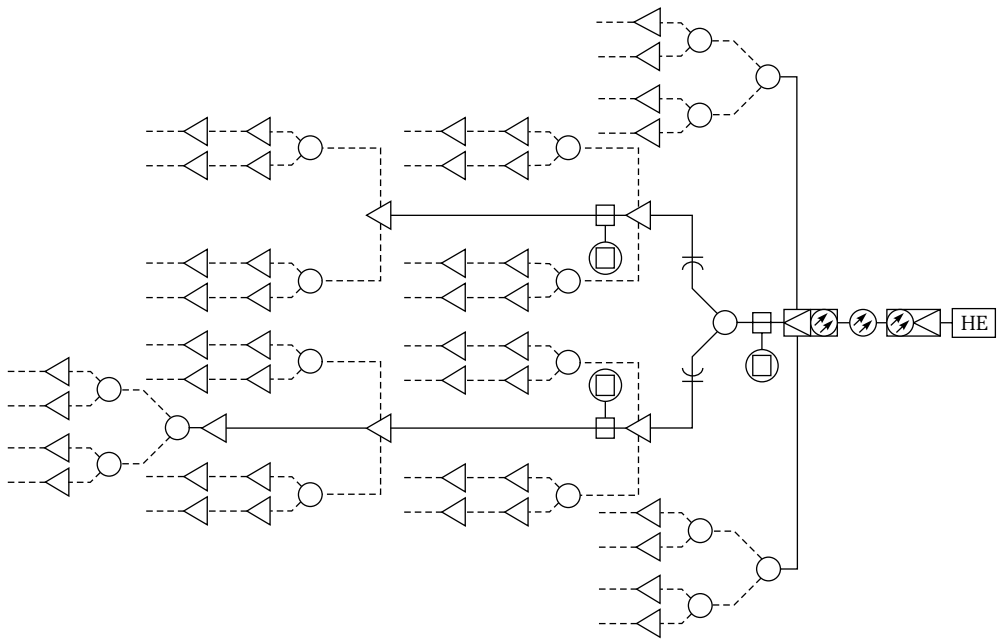


Figure 20.3 Simplified schematic diagram of analyzed system.

The coaxial amplifier cascade beyond the node is limited to 4, and the entire node distribution system contains 53 amplifiers. Three power supplies are required to power all the active devices, with a maximum power supply cascade of two. The total number of series-connected taps in any one distribution leg is about 20.

The initial analysis is based on the use of a generator, but no uninterruptible power supply (UPS), at the headend and field standby power supplies with 2-hour battery capacity. It is assumed that this results in a 30-second headend outage every time the commercial power fails (three times per year) until the generator "kicks in," and that the field standby power supplies have the effect of reducing the field failure rates to 50% per year in each location, with 1 hour of unprotected outage when the batteries do run down (based on dispatching a crew with portable generators as a result of customer-reported outages). It is assumed that there is no status monitoring of power supplies that would have allowed crews with portable generators to back-power supplies before the batteries expire.

20.7.2 Video Services Performance

The network was analyzed for delivery of conventional cable television video services. It was assumed that the viewer was "exposed" to the reliability of 30 pieces of headend channel-processing equipment plus 3 headend combining amplifiers during a typical viewing session, as discussed in Section 20.6.5.

A typical home drop system was assumed, including a set top terminal. No monitoring of the home equipment was assumed so that subscribers were exposed to all failures occurring beyond the tap (since they wouldn't be known to the network operator until reported by the user), but only a portion of the failures earlier in the network, as discussed earlier under user-perceived reliability factors.

The outage rate per year for the most affected customer is 6.11. Though there are no specific standards for telephony service, this is close to CableLabs' suggested maximum of 0.6 per month for video service. (Even though the CableLabs' guideline did not include drop and premises elements, this analysis does since subscriber satisfaction was measured on overall outage rate independent of cause.) Figure 20.4 shows the relative contribution of various network sections to this rate. As can be seen, powering is by far the largest contributor though headend outages also have a major effect than for unavailability. The large power contribution reflects the use of a generator, but no UPS at the headend, so that customers experience a short outage each time the power fails at the headend.

The network availability to the most affected customer was calculated to be 0.99951, equivalent to yearly service outage time of about 4.3 hours. This unavailability is six times better than CableLabs' initial goal for cable television service.

As Figure 20.5 shows, the largest contributor to unavailability is also powering, followed by the drop, headend, coaxial distribution network, and fiber trunking. The large contribution from powering reflects both the limitations of

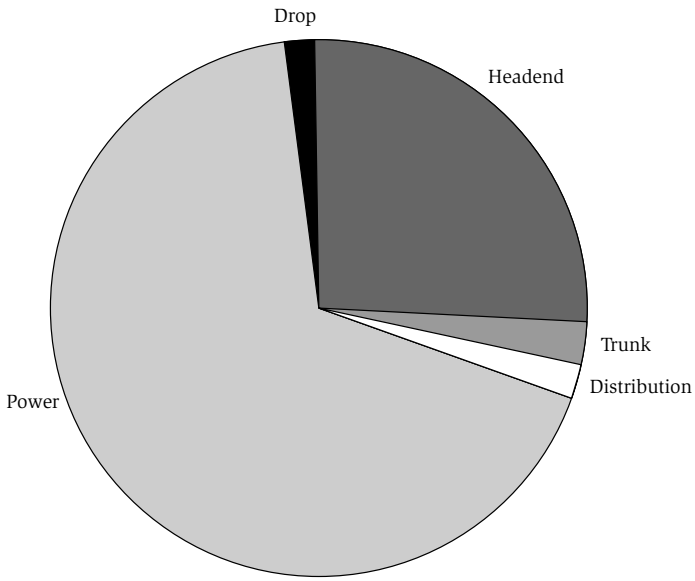


Figure 20.4 Relative network section contributions to outage rate.

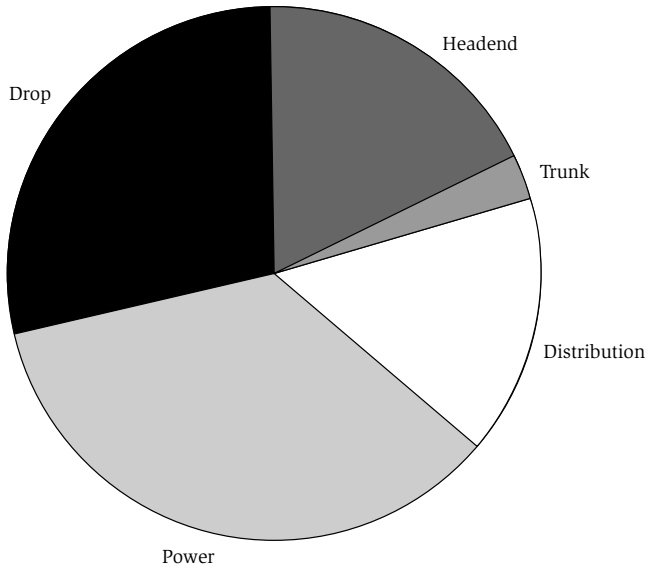


Figure 20.5 Relative network section contributions to outage hours.

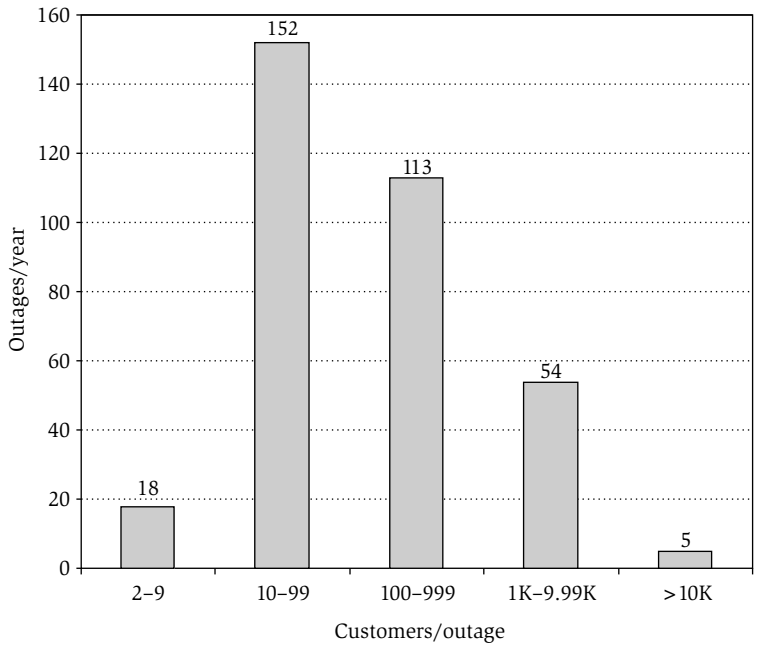


Figure 20.6 Outage distribution by number of affected customers.

battery capacity and the lack of status monitoring. (In Section 20.7.4, we will see the effect of adding power supply status monitoring.) Although the drop hardware is reasonably reliable, its contribution to total outage time is disproportionately large since the MTTR is much higher than for outages affecting multiple customers. The size of the headend contribution is affected primarily by the large number of moderately reliable units involved in producing a multichannel video lineup. Nevertheless, the headend contribution is proportionately less than for outage rate since the MTTR is much less.

A third useful result of the failure analysis is the distribution of multisubscriber outages on the basis of the number of customers affected. This analysis is important because it allows the network operator to size his or her customer service operation knowing the expected volume of outage-related customer calls. Figure 20.6 shows the outage size distribution for the network analyzed.

20.7.3 Wired Telephony Services Performance

The network was also analyzed as a delivery means for wired telephony. In that case, there was no exposure to headend video equipment, and in accordance with

Bellcore methodology, the switch reliability was not included though the interface between switch and distribution system (the host digital terminal) was. In the case of telephony, the drop system included only the drop itself, plus a network interface device (NID), on the assumption that the in-house wiring and analog instruments should not be included in accordance with Bellcore practice. The failure rates assigned to the HDT and NID resulted in 10 and 26 minutes per year of downtime in accordance with Bellcore's allocations in TR-NWT-000418.[24]

When analyzed in this telephony configuration, the outage time is 26% less than for video services, resulting in an availability of 0.999636, whereas the number of outages is 27% less (4.82 per year). Analysis of the causes of the remaining outages shows that power-related issues dominate all others. Although the unavailability is about 3.6 times worse than Bellcore standards for wired telephone service, they are not comparable because of the factors not included in the Bellcore outage definition (this analysis, for instance, includes powering and measures outages from actual occurrence, not from when they are reported).

20.7.4 Effect of Improved Powering on Telephony Performance

If a UPS is added to the headend and status monitoring to the field supplies (so that crews can deliver temporary generators before the batteries lose their charge), then all power-related outages can be eliminated except those due to the reliability of the powering equipment itself. In that case, the telephone service availability improves to 0.99976, and the outage rate declines dramatically to 0.82 per year.[25] Having said that, however, it must be pointed out that the cable industry's experience with early status-monitoring equipment was often that it was less reliable than the equipment being monitored.

As Figure 20.7 shows, powering is still a significant contributor to network unavailability but is less important relative to the drop (dominated by the NID failure rate and slow repair time) and the distribution network. Figure 20.8 shows that the number of large outages has dropped dramatically as a result of improved powering.

It is clear from this analysis that achieving a true 0.9999 availability to the consumer, should that be required, requires several fundamental changes in this network:

- Improved HDT and NID reliability — their Bellcore-assigned outage time allocations consume 68% of the 53-minute total allowed.
- Hardened and more reliable powering. The field power supplies installed by most telcos for their HFC networks were extremely reliable (and expensive) by traditional cable television standards.
- Shorter cascades of both coaxial equipment and power supplies. In general, telco-rated HFC networks are powered by a single power supply located at the node and have active equipment cascades beyond the node of only two to three.

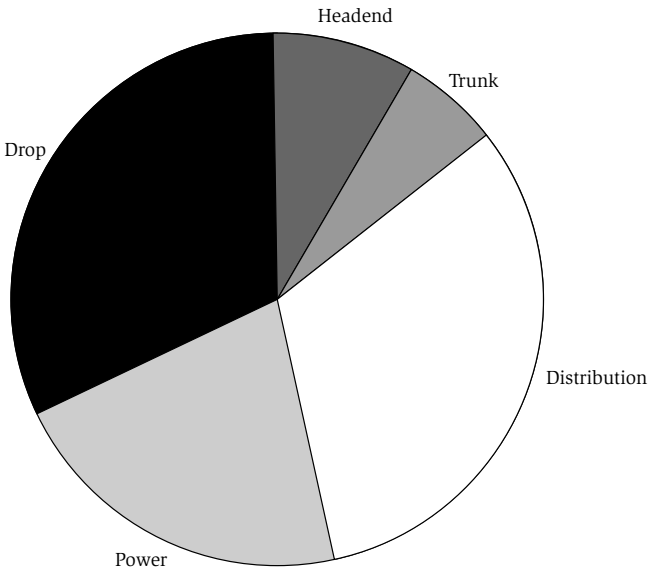


Figure 20.7 Reduced effect on unavailability of powering with headend UPS and status monitoring.

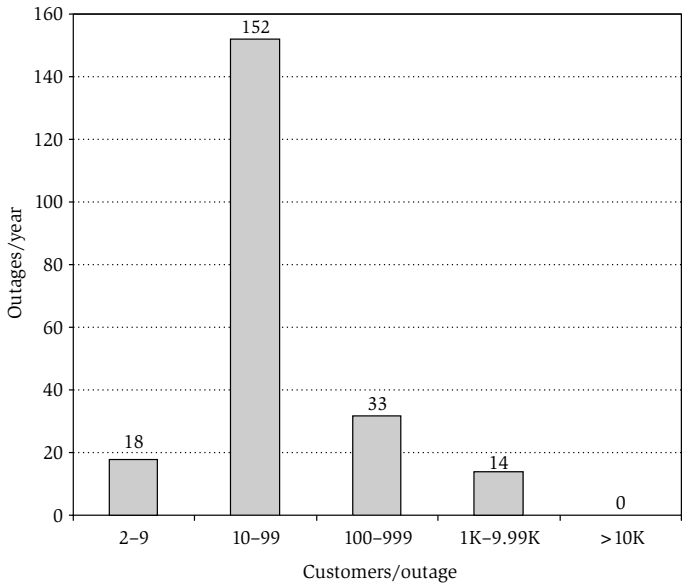


Figure 20.8 Reduction in number of large outages with improved powering.

- Reliable status monitoring throughout the network. Since the time required to be aware of a failure, then to analyze its cause, is a material part of total MTTR, extensive status monitoring plays an essential part of reducing repair times.

Note that the fiber trunking portion of the network was a minor contributor to unavailability in all cases, even in this simple star network. Thus, use of various ring topologies will reduce the instance of large outages (sometimes an important consideration) but will not significantly reduce network unavailability in this network. As other factors are brought under control, the fiber contribution will become more important.

20.8 Summary

The estimation of network availability, failure rate, and other reliability-related factors is very straightforward, if somewhat tedious, given a topology, known component failure rates, and repair times. Unfortunately, a great deal of the effort needed to produce accurate reliability predictions is involved in developing believable component and commercial power failure rates. Once a logical model of a proposed network is constructed and entered into a spreadsheet for analysis, it is simple to test the effect of different component choices and operating practices.

The preliminary analysis done in this chapter illustrates that modern HFC networks are easily capable of achieving low customer-experienced outage rates and unavailability for video services.

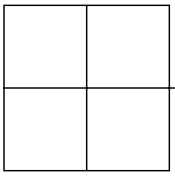
HFC networks that are designed properly for wired telephone services are capable of achieving the required availability, whereas lower-cost networks may be entirely adequate for various levels of video and data services. It is the function of the network engineer to design the most cost-effective network to carry the desired services.

Endnotes

1. Bellcore TA-NWT-00909, Issue 2, Generic Requirements for FITL Systems Availability and Reliability Requirements, December 1993, p. 13-1. Bell Communications Research, Piscataway, NJ.
2. Brian Bauer, In-Home Cabling for Digital Services: Future-Proofing Signal Quality and Minimizing Signal Outages, *1995 SCTE Conference on Emerging Technologies*, Orlando, FL, pp. 95–100. SCTE, Exton, PA.
3. Bellcore TR-NWT-000418, Issue 2, Reliability Assurance for Fiber Optic Systems System Reliability and Service Availability, December 1992, p. 31. Bell Communications Research, Piscataway, NJ.

4. Network Reliability Council, *Final Report to the Federal Communications Commission, Reliability Issues — Changing Technologies Working Group, New Wireline Access Technologies Subteam Final Report*, February 1996, p. 13.
5. CableLabs, *Outage Reduction*, September 1992. CableLabs, Louisville, CO.
6. Ibid., p. II-5.
7. Ibid., p. V-10.
8. Network Reliability Council, op. cit., p. 15.
9. Private communication from Andrew Large. The formulas are valid for cases where $T_c + MTTR < T_s$.
10. This statement is not true when considering *signal* as opposed to *hardware* reliability in the upstream direction since ingress and noise from all branches will affect the reliability of communications on any given branch. Also, as will be seen, we are talking about logical, not physical, cascade here, and the powering configuration may require the modifications to what we would normally consider the cascade order. Finally, electrical short circuits in network branches that are not part of the signal cascade can nevertheless affect communications.
11. Network Reliability Council, op. cit., p. 16. The Figures are quoted from the paper authored by Allen L. Black and James L. Spencer entitled, An Assessment of Commercial AC Power Quality: A Fiber-in-the-Loop Perspective, and published in the *Intelec '93 Proceedings*.
12. Tony Werner and Pete Gatseos, Ph.D., Network Availability Consumer Expectations, Plant Requirements and Capabilities of HFC Networks, *1996 NCTA Technical Papers*. NCTA, Washington, DC, pp. 313–324.
13. Chuck Merk and Walt Srode, Reliability of CATV Broadband Distribution Networks for Telephony Applications — Is It Good Enough?, *1995 NCTA Technical Papers*. NCTA, Washington, DC, pp. 93–107.
14. Nick Hamilton-Piercy and Robb Balsdon, Network Availability and Reliability. *Communications Technology*, July 1994, pp. 42–47.
15. Network Reliability Council, *Network Reliability: A Report to the Nation, Compendium of Technical Papers*. National Engineering Consortium, Chicago, June 1993, pp. 1–32.
16. Bellcore TA-NWT-00909, Issue 2, Generic Requirements for FITL Systems Availability and Reliability Requirements, December 1993, p. 13-5-6. Bell Communications Research, Piscataway, NJ.
17. Tony Werner and Pete Gatseos, op. cit., p. 318.
18. CableLabs, op. cit., p. V-16.
19. Network Reliability Council, *Network Reliability: A Report to the Nation, Compendium of Technical Papers*, National Engineering Consortium, Chicago, June 1993, p. 22.

20. One large MSO reports systemwide MTTRs ranging from 1.64 to 2.39 hours for multiple customer outages. The measured systems are all HFC with various sizes of nodes and wide deployment of both fiber node and coaxial amplifier status monitoring. All headends have UPS plus generator backup. Private correspondence from Nick Hamilton-Piercy.
21. David Large, User-Perceived Availability of Hybrid Fiber-Coax Networks, *1995 NCTA Technical Papers*. NCTA, Washington, DC, pp. 61–71.
22. David Large, Reliability Model for Hybrid Fiber-Coax Systems, *Symposium Record, 19th International Television Symposium and Exhibition*. Montreux, Switzerland, June 1995, pp. 860–875.
23. Nick Hamilton-Piercy and Robb Balodon, op. cit.
24. Bellcore TR-NWT-000418, Table 4, Loop Downtime Allocations — FITL Systems, December 1992, p. 37. Bell Communications Research, Piscataway, NJ.
25. The average monthly video availability among cable systems comprising one large MSO (that has widely deployed status monitoring) varied from 0.9992 to 0.9996 over a 10-month period. Private correspondence from Nick Hamilton-Piercy.



Chapter 21
Analog Video Reception

21.1 Introduction

To a large degree, the modern cable television industry is founded on the ability to deliver premium programming for which subscribers pay, either on a monthly subscription basis or on a per-program basis. Key to offering premium programming is the ability to deny the program to anyone who has not paid for it. The denial technology must offer an adequate level of program denial combined with a high-quality recovered signal and reasonable cost. It must be transparent to the paying subscriber, and it must offer reasonable immunity to defeat. These requirements conflict and have led to many advances in technology.

In the analog era, the cable TV industry depended on equipment manufacturers to provide program denial systems, which resulted in incompatible proprietary scrambling and communications systems. As the industry entered the digital era, it developed sets of standards to achieve a degree of interoperability between equipment made by different manufacturers. Theft of premium services has always been a serious problem in analog transmission, and the hope is that the standards, besides promoting interoperability, will mitigate the theft problem. In this section we shall discuss analog and digital program denial technologies used in the industry. However, we shall stop short of describing any technologies that might be used to circumvent scrambling systems.

The word *scrambling* is used to mean modification of the signal such that it cannot be viewed on a television receiver. A *descrambler*, which has some unique piece of information, is able to convert the signal back into a form visible on a television set. This usage of the word *scrambling* is contrasted to the unrelated usage in the data transmission field. In data transmission, the word refers to the process of modifying a datastream to remove any repetitive bit patterns for the purposes of equalizing the spectral density of the signal and aiding in data recovery. It is common in the data communications field to provide for separate access denial, the technology of which is called *encryption*.

We use the word *pirate* to refer to anyone who attempts to receive programming for which no payment has been made. Usually, it is applied to those who either develop or sell the technology to receive such programming. Sometimes,

though, the word is used to refer to a subscriber who purchases equipment to receive unauthorized programming.

21.2 Non-Set-Top Resident Program Denial

The first class of program denial technologies to be discussed are those that normally don't reside in set top terminals. These have also been known as *whole-house* systems because they operate on signals before they enter the subscriber's home. A signal enabled with any of these technologies is available to all TV sets in the home that are connected to cable.

21.2.1 Negative Traps

The first technology widely used for denial, and one that still has adherents, is called *negative trapping*. Signals are sent over the cable in the clear, and a trap is installed at each home that does *not* take the program. Negative traps are frequently employed for a service with a high take rate, that is, with a large percentage of subscribers who buy a subscription. By protecting such services with a negative trap, the operator needs to supply a trap only for the minority of customers who don't take that service.

Figure 21.1 illustrates the functioning of a negative trap. The filter has a notch, or trap, at the picture carrier of the channel to be denied. By removing the picture carrier (and lower-frequency sidebands), the TV has no signal to recover. Since no signal enters the home, the technology is rather secure. The biggest security threat is unauthorized removal of the trap from a drop. This may be discovered by physical audit in many cases though in some apartment complexes auditing may be a problem. The other issue in security is that the trap may fail, allowing signals to pass. Traps may drift with time or temperature so that the notch moves off the picture carrier. Enough signal power is then delivered that the TV can recover a picture, although a degraded one.

A disadvantage of a negative trap is that it will often have an effect on the lower adjacent channel. The frequency to be removed is only 1.25 MHz away from the channel edge and 1.5 MHz above the adjacent channel sound carrier. If the trap is used on channel 3, the adjacent sound carrier is separated from the notch frequency by 2.4%, so it is relatively easy to control the rise of the notch. However, if the notch is on channel 13, the adjacent channel sound is separated only by 0.7%, making the notch much more difficult to realize. The nature of a filter is that the transition band is proportional to frequency all else being equal. Also, temperature drift is a constant percentage of the filter frequency. Thus, traps are normally restricted to use in the lower portion of the cable spectrum. As you go to higher frequencies, the effectiveness of the notch decreases, and the problems it causes increase.

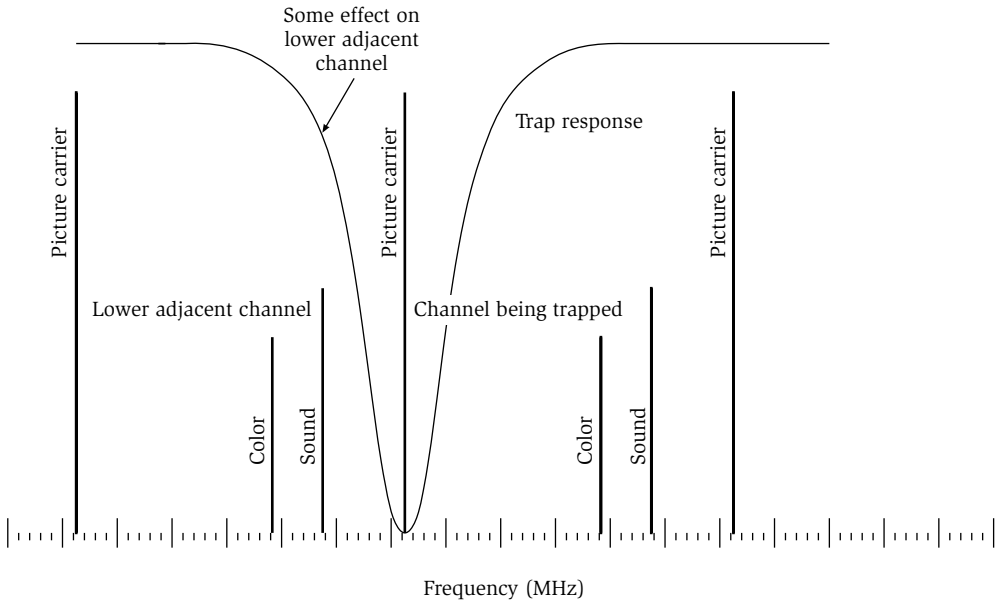


Figure 21.1 Operation of a negative trap.

A trap will also introduce group delay in the lower adjacent channel. Group delay is introduced near the transition region of a filter: signal components closer to the band edge of a filter are delayed with respect to those further removed. The lower adjacent color carrier is closer to the transition region than is the lower adjacent picture carrier so the color information will be delayed with respect to the luminance information. If the group delay is large enough, it produces the so-called funny paper effect, in which color information is shifted with respect to the black-and-white information in the picture.

21.2.2 Positive Traps

A similar technology is that of positive trapping. The shape of a positive trap is nearly identical to that of a negative trap, but the trap frequency and method of use are different. An interfering carrier is inserted at the headend, and a trap is used to remove the interfering carrier at the home of a subscriber who takes the service. Positive traps are popular for services having low penetration since they are installed only at the homes of the people taking the service.

Figure 21.2 illustrates the operation of a positive trap. One or two jamming carriers are inserted at the headend, usually 2.25 MHz above the picture carrier. The peak amplitude is usually the same as, or slightly higher than, the picture

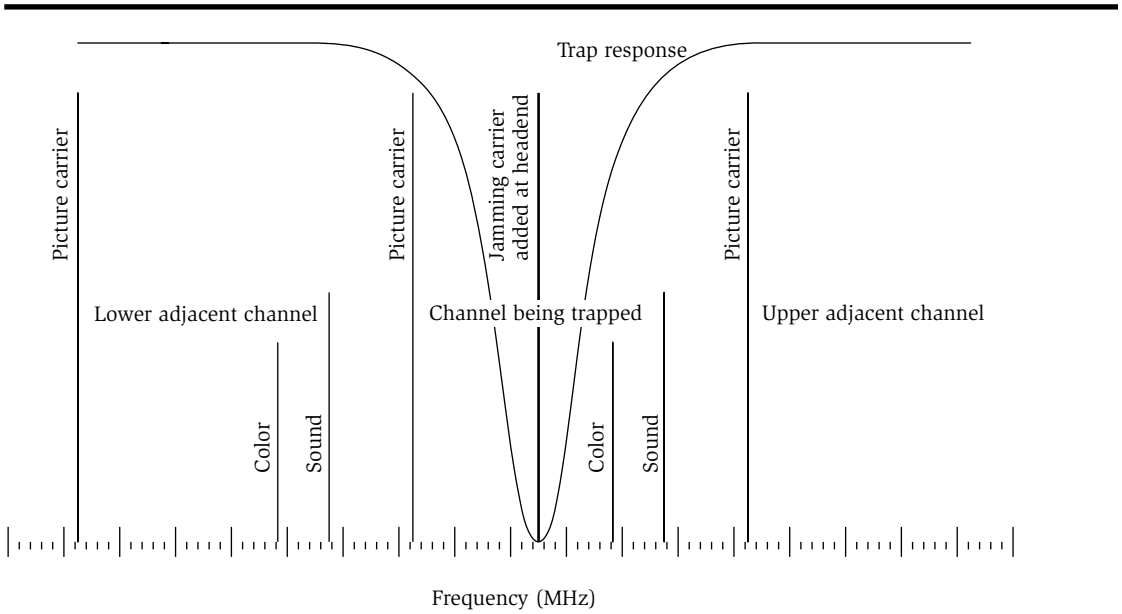


Figure 21.2 Operation of a positive trap.

carrier amplitude. The jamming carrier(s) may be modulated with various signals. In a mixing process, such as the picture demodulator in the TV, the second harmonic of the jammer (with respect to the picture carrier) is generated and falls on the sound carrier, affording a large degree of scrambling for the sound carrier.

The positive trap is located at the jamming carrier frequency so as to remove the jammer. Assuming the jammer is completely removed, the TV will be able to recover the signal. Since the trap cannot be made infinitely narrow, it will remove some desired modulating sidebands as well as the jamming carrier. Except at the jamming frequency, this may be at least partially compensated for by preemphasizing the midband frequencies at the headend. Preemphasis accentuates the frequencies that will be partially removed later such that the net result is an approximation of normal sideband levels. Preemphasis may also be applied to the group delay introduced by the trap. Because the amount of power around 2.25 MHz in a typical video signal is low, the picture damage is often considered acceptable. Nonetheless, positive traps have a reputation for "softening" a picture.

Since the positive trap notch frequency is toward the center of the channel, we would expect less effect on the lower adjacent channel than that caused by a negative trap. Since the trap can be located inside the subscriber's home, the temperature range over which the trap must function is reduced.

A jamming carrier generator is connected to each modulator carrying a signal to be positively trapped. This jamming carrier usually interfaces with the modulator using either the picture IF loop or the composite IF loop of Figure 8.3. In addition, the baseband video signal may be predistorted to compensate for the amplitude and delay effects described earlier. This can be done either at baseband or at IF.

Engineers must be cautious when measuring the level of a signal that includes a jamming carrier. Depending on the instrument used to measure the level, the jamming carrier may or may not interfere with the measurement. The easiest way to determine whether or not a measurement is being thrown off by the jamming carrier is to turn off the jammer while measuring the signal level. If the measured level changes, then the jammer was interfering with the measurement.

A variation on positive trap technology is a system that replaces the jamming carrier with a highly emphasized copy of the visual signal: an emphasis circuit at the headend amplifies the baseband video spectrum around 2.25 MHz so much that the power at that frequency is about as great as at the picture carrier, so it forms its own jamming signal. This "jamming signal" is restored to more-or-less normal level by the positive trap at the home.

21.2.3 Interdiction

Interdiction is a program denial technology that can offer addressable services and good security, along with the ability to provide whole-house service. Signals are transmitted in the clear on the cable plant. The normal tap is replaced with a special tap that includes jamming oscillators whose frequencies are close to that of the picture carriers to be denied. The jamming is applied only to those channels that the subscriber has not purchased. Since the jammer is not intended to be removed, it can be located arbitrarily close to the picture carrier, where removal is more difficult. One system actually allows the jamming carrier to wander from one side of the picture carrier to the other.

Figure 21.3 represents a generic interdiction system, which takes the place of the normal tap. The interdiction system includes a directional coupler DC1 and an amplifier A1. After amplification, the signal is split in CM1 so as to be routed to all tap ports. Directional coupler DC2 routes signals to the data receiver or transceiver of the addressing system. Directional coupler DC1 and splitter CM1 compose a normal tap. (In this text, splitters are referred to using the designation CM for combiner because the same device is used for either task.) Each port has another amplifier, used mainly to provide isolation for the jamming carriers, to prevent them from leaking to other outputs. A series of jamming oscillators is provided, which can be switched in and out under command of the addressing system. When an oscillator is connected to a port, it is set to a frequency close to, and sometimes equal to, the picture carrier. The signal received in the home is rendered useless because of the jamming.

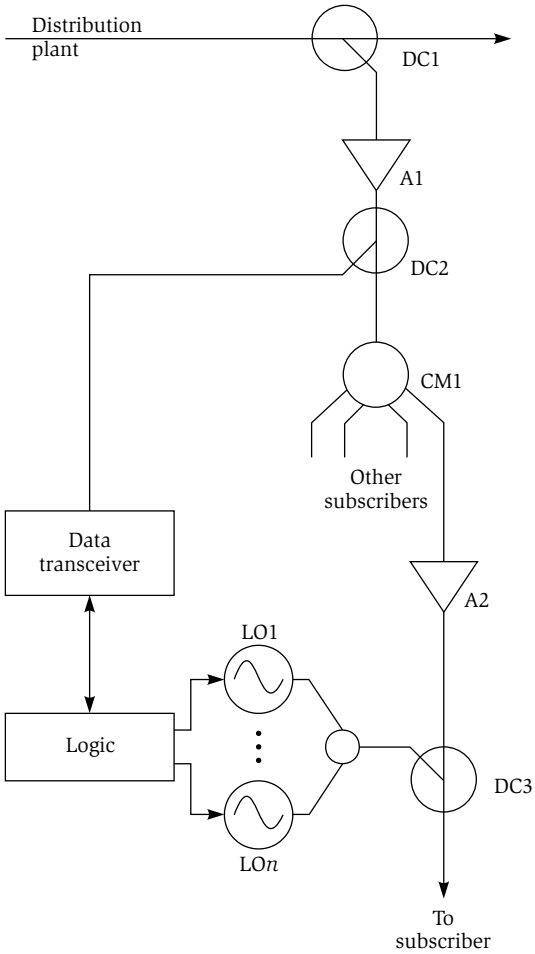


Figure 21.3 Interdiction system.

The advantage of the interdiction system is the ability to control a number of channels, allowing the operator to provide advanced programming options such as can be provided using a set top terminal. At the same time, interdiction provides excellent security with no equipment in the home. It does not interfere with the use of VCRs or picture in picture, and the signals are available to all TVs in the house.

The disadvantages include the need to supply equipment for every subscriber, regardless of his or her purchase of premium services. Powering is also an issue: the equipment must be mounted away from a subscriber's reach so that he

or she doesn't have access to the unscrambled signals, which exist on the plant side of the interdiction equipment. This normally means mounting at the tap location. Powering is from the plant, so the cable operator assumes the power bill rather than its being assumed by the subscriber.

21.3 Set-Top Resident Program Denial—Analog Modulation

The dominant program denial methods involve use of set-top terminals (STTs) in the home. Although many technologies have been proposed to effect scrambling, only two are in common use in North America: sync suppression and video inversion. The security afforded with these program denial methods is not total. However, the industry has decided, at least for analog signals, that other methods are not cost effective. Indeed, a review of the plethora of attacks on signal security in the last 20 years reveals that the most successful attacks against reasonably modern scrambling methods are those commonly called "spoofing." In spoofing systems, a modified set top terminal is "spoofed" into "believing" that it is authorized to descramble a particular signal. The attack is through the authorization system, not the scrambling system. Though there have been attacks against the scrambling system, they usually either have been minimally successful or have yielded to signal modifications in the headend.

21.3.1 Sync Suppression Scrambling

The most common scrambling method in North America is, by far, sync suppression. Variants of sync suppression have been employed since the earliest practical scrambling systems, dating back at least to the mid-1970s.

Recall from Section 8.5 that the sync tip corresponds to the highest amplitude of the modulated RF envelope. If the sync is somehow modified by moving it to amplitudes occupied by active video information, then a television set will not be able to synchronize. The effect is a screen full of random, more-or-less horizontal, lines.

The Scrambling Process

In order to illustrate certain features of scrambling systems, a reference waveform is used, as shown in Figure 21.4. Scrambling operations will be illustrated with reference to this waveform. (Refer to Chapter 2 for more information concerning the video waveform.) The reference waveform shown in Figure 21.4 is a ramp, which rises from black, at the left side of the screen, to white, at the right. The setup shown is a feature of NTSC video that is left over from the early days of television. Color burst is shown though no color information is shown during the active line. In general, we would expect additional color subcarrier power during the active line.

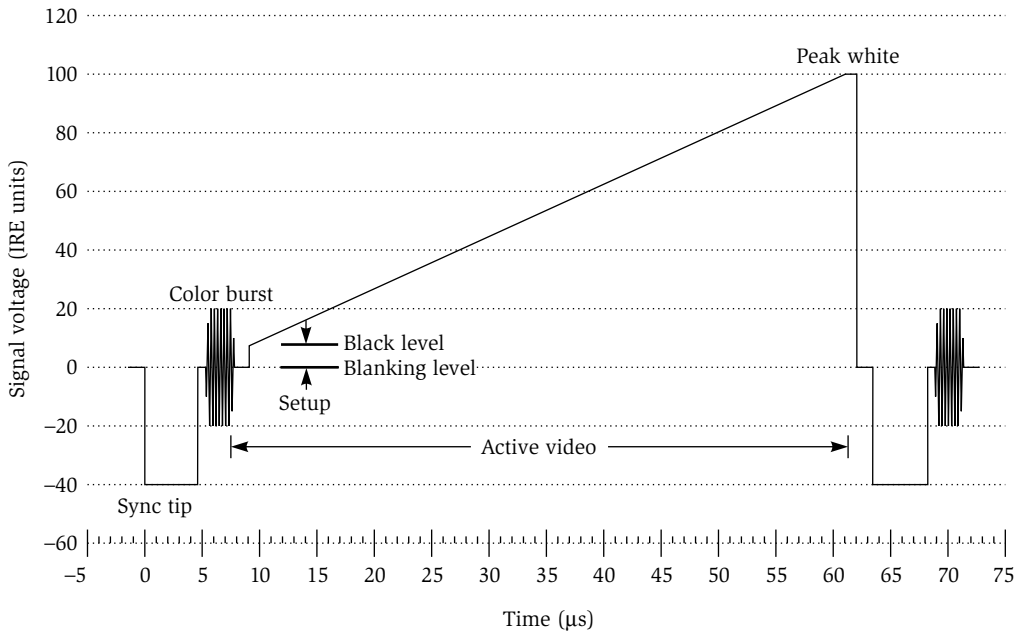


Figure 21.4 Reference video waveform used to illustrate scrambling systems.

Figure 21.5 illustrates a typical relationship between a modulator and an RF sync suppression scrambler. The portions of the modulator shown are identified as they are in Figures 8.3 and 8.4. Flags B and D are as shown in those two figures, where they were used to route signals from one figure to the other. Baseband video is looped through the scrambler, where a high-input impedance amplifier couples video to a sync separator. The scrambler must have access to the video sync so that it can synchronize the suppression and restoration information. The process of routing video through a piece of equipment is often called *loop-through* and was described in Section 8.9.1.

From the loop-through input of the scrambler, the video is routed to the modulator, where it is processed as described in Chapter 8. In the scrambler, the sync separator extracts vertical and horizontal sync, which are passed to timing circuitry. The timing circuitry operates a switched attenuator, AT10, in the visual IF path, and an amplitude modulator, M10, in the aural IF path.

The top left of Figure 21.5 shows the reference waveform of Figure 21.4 modulated onto the IF carrier. As explained in Chapter 2, the modulation format prescribed for NTSC (and PAL) video is negative modulation, in which the sync tips correspond to the highest amplitude of the RF (or IF) signal, and peak white corresponds to the minimum amplitude. Sync can be suppressed simply by

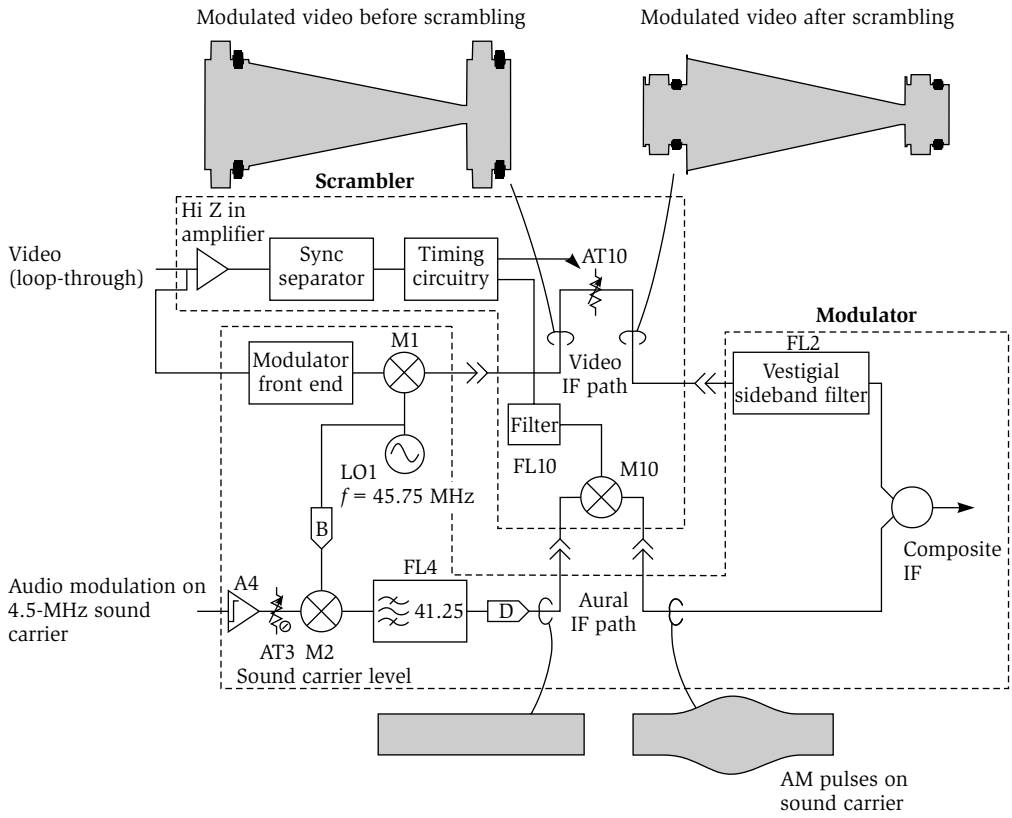


Figure 21.5 Sync suppression scrambling circuitry and waveforms.

switching in an attenuator, AT10, during the sync time. This results in the waveform shown at the top right of the figure. During the horizontal (and, usually, vertical) blanking intervals, AT10 is switched in, and during the active video it is switched out. Now the sync tips no longer correspond to the highest amplitude of the modulated waveform. Two values of attenuation are commonly employed: 6 and 10 dB. The waveform shown has been scrambled using a 6-dB attenuator. Had a 10-dB attenuator been used, the horizontal blanking interval would be of even lower amplitude.

From attenuator AT10, the signal is returned to the modulator, where it passes through the vestigial sideband filter FL2. The process of switching AT10 effectively applies amplitude modulation to the visual IF. This modulation produces sidebands that would extend outside the television channel were they not filtered in FL2. After the vestigial sideband filter, the signal is combined with the aural IF, with the remaining signal processing being as shown in Figure 8.4.

Inspection of the scrambled video waveform at the top right of Figure 21.5 shows that the peak amplitude of the scrambled signal is not constant. The peak amplitude depends on the darkest portion of the picture. Because of this, it is not possible to use a sync-suppressed signal as a pilot to operate the AGC of a coaxial or microwave distribution link. Furthermore, it is not straightforward to measure the signal level of a sync-suppressed signal. Manufacturers normally recommend turning off scrambling before measuring signal level. The manufacturer must then ensure that the scrambled signal level remains correct when scrambling is turned on.

Some sync suppression systems take advantage of the reduced peak amplitude of a sync-suppressed signal to raise the signal level about 2.5 dB when scrambling. This returns the peak amplitude of the scrambled signal to the peak amplitude of the nonscrambled signal, thus improving carrier-to-noise ratio by 2.5 dB during active video.

Synchronizing the Descrambler

Several methods are used to communicate descrambling information to the descrambler. One of the more common is to place pulses on the sound carrier, as shown in the lower portion of Figure 21.5. The aural IF is routed out to the scrambler at flag D. Amplitude pulses are modulated onto the sound carrier as shown by the envelopes in the bottom of the figure, where the left envelope is unmodulated, and the right envelope has the modulation impressed on it.

Recall that the audio is frequency modulated onto the sound carrier. Theoretically, the carrier may be amplitude modulated with an independent signal, such as the scrambling synchronization waveform shown, and the two will not interfere with each other. In practice, a number of mechanisms can cause crosstalk between the amplitude and frequency modulation components. The manufacturer must take pains to control these sources of so-called AM/FM and FM/AM conversion. It is possible, though not common, that the distribution plant can introduce undesirable AM/FM or FM/AM conversion.

The sound carrier occupies a relatively narrow spectrum (see, for example, Figure 8.2). You must carefully control the spectrum that results from impression of the AM modulation on the carrier. For this reason, filter FL10 is used in the path between the timing circuitry and modulator M10. The filter controls the spectrum of the modulat*ing* signal and, hence, of the modulate*d* signal. It reduces the rise and fall time of the modulating waveform, resulting in the slow transitions shown in the right envelope. A fast transition would be preferable to allow accurate recovery of timing information from the waveform because noise can distort the recovered waveform. The slow rise time means that such distortion could translate into timing errors. However, fast rise times mean the spectrum will spread, possibly causing interference with either the on-channel or adjacent channel video. The manufacturer must walk a fine line between timing recovery accuracy and spectrum spreading.

Descrambling is effected with an attenuator in the descrambler that operates backward from attenuator AT10: the descrambler attenuator must be IN during active video, and OUT during blanking intervals when AT10 is IN. This operation must be timed very carefully to coordinate the operation of the two attenuators to within tenths of microseconds. Besides adjusting the value of attenuator AT10 to effect variations of scrambling in order to frustrate pirates, some systems also vary the timing of the aural carrier pulses with respect to the video.

A variant of the scrambling system described here is baseband sync suppression scrambling. The sync is suppressed before the video is modulated. Because of the way the systems most logically work, baseband sync suppression is accomplished by offsetting the sync level rather than multiplying it. Baseband sync suppression systems don't require the visual IF interface shown, and since commercial systems don't rely on pulses on the sound carrier, they don't use the aural IF interface either. However, because the sync is not the most negative portion of the waveform as it enters the modulator, the clamp in the modulator (see Figure 8.4) has to be modified to clamp at the modified sync tip level, and a signal must be supplied to the modulator to define the clamp time.

Characteristics of the Scrambled Waveform

Figure 21.6 illustrates the demodulated waveform that would be processed by the baseband circuits of a TV receiving the scrambled signal. Also shown, for reference, is the nonscrambled waveform in dashed lines. Note that the sync, which should be the most negative part of the waveform, has been moved up to a voltage range normally reserved for picture information. The darkest portion of the picture is now the most negative portion of the waveform, assuming the picture is not all "whiter" than the modified sync tip level. Sync circuitry in the TV will try to synchronize to the wrong place in the line.

As illustrated, the most negative portion of the waveform occurs at both the beginning and end of the line, where the picture content was black as scrambling ended and started. With most real video patterns, the most negative portion will be somewhere else along the line and will vary from one line to another.

Notice that the peak-to-peak amplitude of the sync tip and the color burst, which is also suppressed, is reduced to one-half that of the unscrambled signal. This is a characteristic of RF sync suppression systems. Since attenuator AT10 of Figure 21.5 reduces the amplitude of the signal by one-half (for 6-dB suppression), all portions of the suppressed signal are reduced in amplitude by one-half. Baseband sync suppression, on the other hand, operates by shifting the signal level during the HBI using a voltage added to the video signal. The peak-to-peak amplitude of the sync and the burst is not reduced with baseband scrambling.

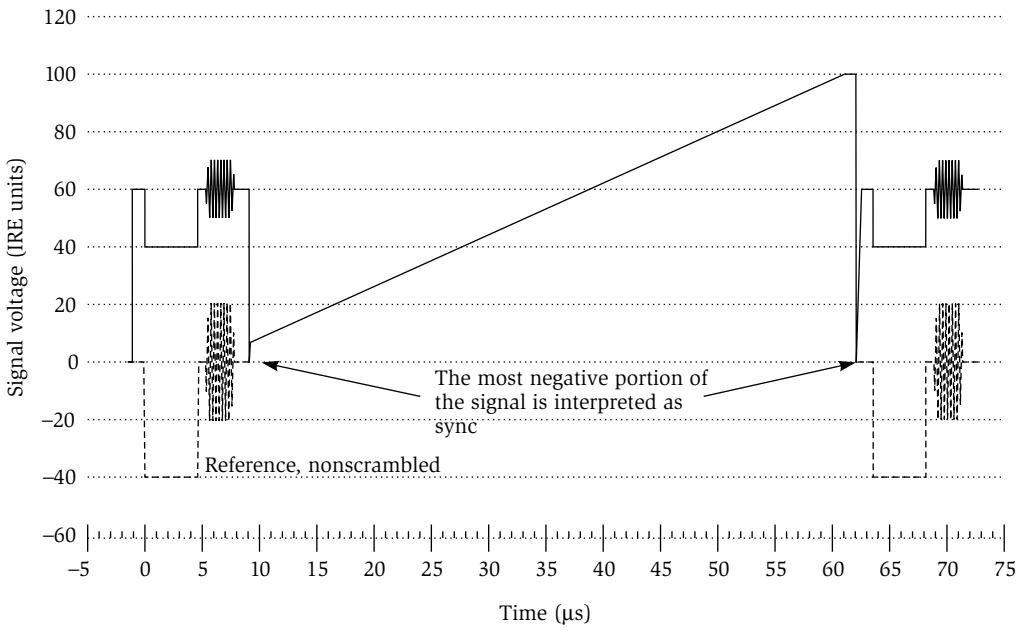


Figure 21.6 Scrambled signal recovered at television receiver.

The Descrambled Waveform

Figure 21.7 illustrates the signal as it would appear in a TV that has received the signal after descrambling. The horizontal (and usually vertical) blanking intervals have been suppressed in the scrambler and expanded in the descrambler. The order of events is very important. If the descrambler ever expanded (by switching out attenuation) a signal that had not been suppressed, the result would be a sliver of signal more negative than sync. This could cause unacceptable operation of the TV receiver.

Because some time must be allowed between suppression and expansion, a small sliver of time remains in the descrambled signal, during which the waveform was suppressed but not expanded. This artifact of the scrambling and descrambling process is characteristic of sync suppression systems, unless removed by another signal-processing step. The artifact is sometimes called a "rabbit ear."

The rabbit ears of Figure 21.7 can have deleterious effects on the signal. One early closed caption decoder required all the run-in pulses transmitted at the beginning of the closed caption line (line 21). In a particular cable system, the scrambler was not set properly, resulting in the rabbit ear at the beginning of line 21 impinging on the first closed caption run-in pulse. The problem was fixed by

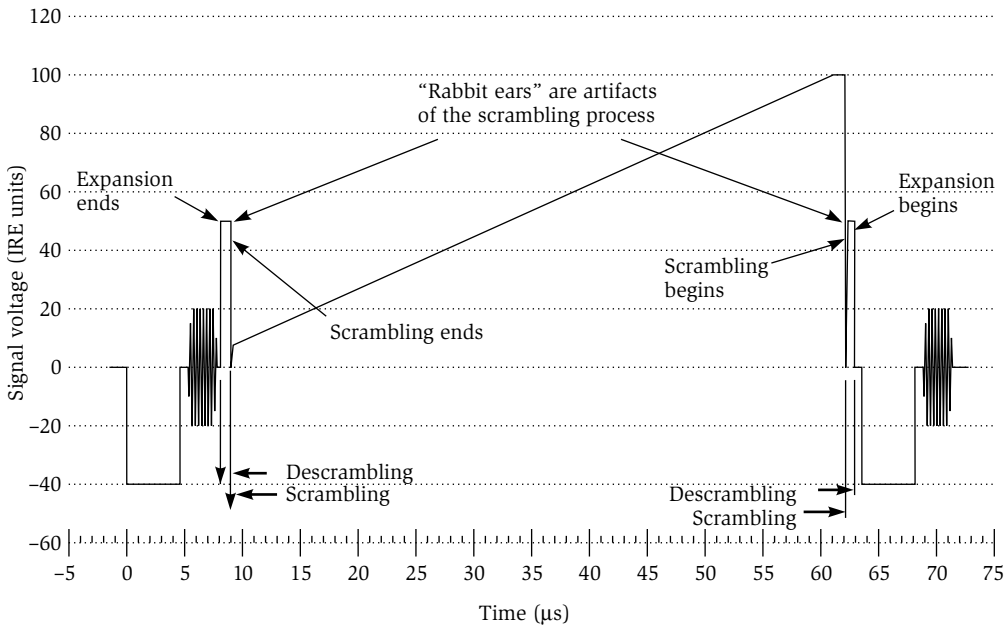


Figure 21.7 Descrambled signal at authorized TV receiver.

moving the timing in the scrambler so that the run-in pulse was spared. (Run-in pulses appear at the beginning of the line of closed caption data and are used to synchronize the decoder.)

The rabbit ears are normally in what is known as the *overscan* region of the TV picture. That is, the electron beam in the TV scans off the edge of the picture tube screen and does not show the rabbit ear. Occasionally, a TV is encountered that does not exhibit normal overscan: the picture does not cover the entire screen plus a little spilling off the edge. In this case, a white bar may be visible on one side of the screen.

Sometimes electrons emitted during the rabbit ear reflect from the side of the picture tube and strike the phosphors on the screen. The electron beam is badly defocused but can result in a slight lightening of the picture on the edges. Usually, the effect is not noticeable because it is always present. However, in scrambling systems that change scrambling modes such that the rabbit ear changes, the effect can be noticeable at a scrambling mode change. For this and other reasons, scrambling modes are normally changed at scene changes.

Yet another issue relating to the rabbit ear is the minority of television sets that perform dc restoration (clamping) on the back porch rather than on the sync

tip. If the set tries to clamp when the rabbit ear occurs, the clamp level will be in error, leaving the picture darker than it should be. A back porch clamp can be identified by changing the timing of the back porch rabbit ear and noting if the picture brightness changes. Such a test is best done using video from a test generator.

The suppression interval is determined solely at the scrambler, which will usually contain some sort of adjustment to set the suppression start and finish times. The expansion interval begins at a time related to the time of a pulse on the sound carrier (or some other time marker depending on how the system works). The end of the expansion interval is then timed from the beginning of expansion. Because of the criticality of the sound carrier pulse timing, it must be set at the scrambler/modulator. Reference to Figure 21.5 will show that the sync-suppressed picture carrier is routed through the vestigial sideband filter FL2. The sound carrier, with the synchronizing pulses, is frequently not routed through FL2 (though some manufacturers may do so).

Modern vestigial sideband (VSB) filters are realized using surface acoustic wave (SAW) technology. A characteristic of this technology is unavoidable delay to the signal traversing the filter. The amount of delay is related to the steepness of the filter transition region. To an extent, it is also related to the present state of the art, but in every practical case, there will be a rather large delay to the signal in the SAW filter. This delay is long, compared with the 100 ns or so timing accuracy desired, making imperative the careful setting of the relative timing of the scrambling and sound carrier pulses. This timing is done in the scrambler, which must be mated with the modulator with which it will be used. If the SAW filter or the scrambler is changed, timing must be reset because a different SAW filter in FL2 may have a different delay.

The scrambler manufacturer will provide a procedure to adjust the scrambler. Usually, the most accurate method involves detecting the video signal and the aural carrier pulses, possibly using a spectrum analyzer, whose output is applied to an oscilloscope. The time between the leading edge of sync and the 50% point of the pulse on the sound carrier is measured and set to a specification supplied by the manufacturer. This technique, though very accurate, is tedious and therefore prone to error. Often a substitute method, using a certified descrambler, is used. This certified descrambler is subject to drift in descrambler timing but is adequate if sufficient care is exercised.

21.3.2 Scrambling by Video Inversion

Many analog scrambling systems used in North America employ video inversion in addition to sync suppression. In video inversion systems, the active video is inverted with respect to its normal polarity. The produces a negative picture image. In addition, since the HBI is rarely inverted, color information is reversed as well: the phase relation between the active color and the burst is changed by 180°.

Figure 21.8 illustrates video inversion of the reference waveform. The nonscrambled waveform is shown as a dotted line. The video in this case is inverted with respect to an axis of 50 IRE. One way to think of the inversion is that the video (but usually not the sync) is rotated about an axis, in this case, of 50 IRE. Mathematically, the operation performed on the active video is given by the following equation:

$$V_{\text{out}} = 2V_{\text{axis}} - V_{\text{in}} \tag{21.1}$$

This axis information, V_{axis}, must be communicated to the descrambler in some manner. If it is not communicated, then gain errors will create an error in the axis about which the video is reinverted (reinversion applies the same equation), resulting in a picture that is too dark or too light. Some early video inversion systems suffered from this problem, but the issue was quickly recognized, and more modern systems include a means for communicating the axis information to the descrambler. One method is to transmit a level in the vertical blanking interval that is related to the axis of inversion. Another method is modification of the sync pulse in some manner to communicate the axis information. Either method can give satisfactory results.

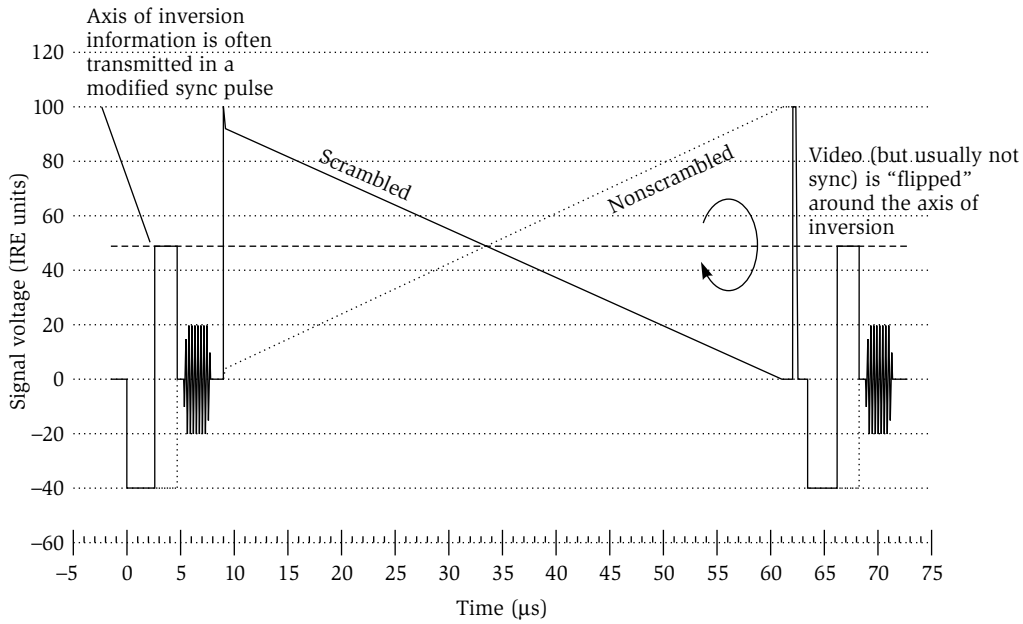


Figure 21.8 Video inversion with the axis of inversion at 50 IRE.

The illustrated scrambled waveform communicates axis of inversion information to the descrambler by shifting the last half of the unsuppressed sync tip to the level of the axis of inversion. Some scrambling systems use an axis of inversion of 50 IRE, and others use 30 IRE or some other axis. Use of 30 IRE would yield a picture in which the white level is translated to the normal unscrambled sync tip level.

Video inversion is often accompanied by sync suppression since inversion by itself doesn't necessarily yield a completely unrecognizable picture. By itself, the effect of video inversion is much like that of looking at a photographic negative. Combined with sync suppression, it makes life more difficult for the pirates. Frequently scrambling manufacturers will combine sync suppression and video inversion, using several variations on each, to yield a number of scrambling modes they can switch between so as to cause more problems for the pirates.

21.3.3 Audio Scrambling

Audio scrambling is used at times as a supplement to video scrambling. In certain cases, cable operators are required to deny sound to subscribers who have not purchased a particular program. Some set-top terminals, by the nature of their operation, don't tune unauthorized channels. Thus, the subscriber never sees a scrambled picture nor hears the audio unless he or she tunes the channel without a set-top terminal. If the subscriber does so, he or she will hear the audio.

The more common audio scrambling methods have involved one of two approaches. One approach is to place monaural audio on the difference channel of a BTSC stereo signal. For a description of BTSC stereo, see Chapter 8. Sometimes a variation of BTSC will be used. A "barker," or teaser, audio is added where the sum signal goes, both to advertise the program and to mask the audio on the difference channel.

Another method used to deny audio is to place a second audio carrier at a frequency other than that normally used for audio (picture carrier plus 4.5 MHz in NTSC transmissions). Normally, the regular aural carrier is also transmitted, with a barker carried thereon. This method allows stereo to be transmitted, but finding the required "clean" spectrum is difficult, and the set-top terminal cost is increased because of the need for a second sound demodulator.

21.4 The Set-Top Terminal

For many years, the set-top terminal (STT) has been a part of subscribers' lives, first as a device that enabled more channels than what a TV could tune. Later, STTs added remote control before that feature was standard in TV sets. STTs are also commonly used as a vehicle for program denial. Most scrambling systems rely on a descrambler built into an STT. A very recent ruling by the FCC requires descrambling to be separated from the STT early in the 21st century.

21.4.1 Set-Top Terminal Classification

STTs are frequently considered to come in three fundamental types. The simplest is the plain converter, which performs no functions except channel tuning and perhaps volume control. This type of converter is the one traditionally used for expanding the tuning of a TV that cannot tune all the channels available on the cable. Though still available, plain converters tend to be sold more at retail today, for use with the dwindling number of TV sets that don't tune all cable channels.

The next class of set top up from the plain set top is the descrambling STT. It is similar to a plain converter, except that it includes some sort of descrambler. Authorization of programs must be embedded in the terminal because the cable operator has no way to communicate with the terminal once it is installed in the home. Older descrambling terminals had programmable read-only memory (PROM) devices installed, which controlled what could and couldn't be descrambled. Later models usually had a facility to be programmed either by plugging a programmer into a socket on the terminal or by using the infrared remote control link. In every case, the terminal manufacturer provided the programming facility. This type of terminal is rarely sold today though many remain in use.

By far, the most popular class of STTs today are addressable terminals, in which the cable operator maintains some sort of communications with the set top when it is in a subscriber's home. This communication can be accomplished in several ways, the most popular of which we describe briefly here. The communications link is used to authorize the terminal to descramble services according to what the subscriber has purchased.

The addressable data sent traditionally has included authorization information for the terminal to specify what channels or programs should be descrambled. Some STTs tune only authorized channels. Others tune all channels and search for program identification information ("tags") to determine descrambling authorization.

Electronic program guides (EPGs) are popular now. In this application, a program schedule is transmitted to the set top terminal, which formats it for display to the subscriber. The subscriber typically moves a cursor on the screen until finding a program he or she wants to view. By pressing a "select" button, the set top is directed to the relevant channel.

Out-of-Band Addressability

Out-of-band addressability uses a separate data carrier to communicate data to the STT. A separate data receiver is contained in the STT. The advantage of out-of-band addressability is that the operator can communicate with the STT at any time. Instructions to change the terminal's state, such as when a pay-per-view event is ordered, can be processed at almost any time, assuming that the terminal is powered and connected to the cable. Consequently, a message does not have to be transmitted much more than one time to ensure receipt.

In-Band Addressability

Out-of-band addressable terminals must bear the expense of a separate data receiver. In-band terminals use addressing information carried in one or more TV channels. This data may be carried in the vertical blanking interval, as additional pulses on the sound carrier, or using some other method. One advantage is the reduction in cost by not having a separate data receiver. Some practitioners feel security is enhanced by using in-band data. The problem usually cited with in-band addressability is the inability to communicate with the set top except when it is tuned to the proper channel(s). To overcome this problem, some manufacturers place data on many different channels, even if the channel is not scrambled. They also creatively divide the data among channels to maximize the probability that a terminal will receive intended data. However, the operator can never be sure that a particular terminal has received all the data sent to it.

Alternative Authorization Procedures

Several methods may be used to authorize programs or channels to be descrambled. This section describes the most common.

Subscription programming is the oldest method of authorizing pay programs. In a subscription system, the subscriber pays a monthly fee and receives a premium channel all the time. This requires that a terminal be addressed only when the subscriber changes the channels to which he or she subscribes. Only one-way communications is required.

Pay-per-view (PPV) programming is sold on a per-program basis, but the subscriber must call the operator in advance to order the program. Ordering may be done by talking to a customer service representative (CSR) with the cable company, who adds the price to the subscriber's bill and causes an authorization signal to be transmitted to the terminal. Alternatively, the customer may call an automated system that uses automatic number identification (ANI, popularly known as caller ID) to identify the caller. The caller punches in a number corresponding to the program desired, and his or her set top is automatically sent an authorization.

Impulse-pay-per-view (IPPV) programming is the preferred method of selling individual broadcast programs since the subscriber doesn't have to preorder: he or she can order a program on impulse, even after it has started. IPPV systems require two-way communications with the terminal because the terminal must inform the headend of programs viewed so as to allow billing and payment for the programs.

Commonly, IPPV systems work in what is called a store-and-forward manner. When the subscriber requests a program by actuating a prescribed key sequence, the terminal allows the program to be viewed without checking first with the headend. (If the headend must provide preauthorization, the resulting

time delay is considered unacceptable.) At a later date, the terminal is polled from the headend and reports what programs have been watched. From the time the program is viewed until the time the terminal reports the viewing, the viewing information is stored in nonvolatile memory. If the subscriber interferes with communications to the headend, the operator will not know to bill for the viewing. To minimize losses in such a case, the terminal limits the number of programs (or the cost thereof) that can be viewed before it communicates with the headend.

Video-on-demand (VOD) programming is perhaps the ultimate in pay programming. VOD systems allow the subscriber to select what he or she wants to see, when he or she wants to see it. Programs are stored on a video file server similar to those used to store local advertising (see Section 8.10). VOD requires the use of RF return signals to allow the subscriber to request a program from some sort of menu. Advanced systems include so-called VCR functions of pause, rewind, and fast forward. For this to work, the RF return path must exhibit low delay (latency) so that the experience is truly similar to that of using a VCR.

Near-video-on-demand (NVOD) is a compromise system used when an adequate number of channels are available. Selected movies are transmitted, each movie on several channels. On each channel used to transmit a particular movie, the start time is staggered by 10–30 minutes. This way, a subscriber gets convenient starting times. By tuning to a channel with a later start time, it is possible to simulate "pausing" the movie.

Subscription video-on-demand (SVOD) allows subscribers to pay a fixed amount per month to have access to a library of programs they can view at any time, as if they were VOD programs. The difference is that, rather than pay per program viewed, the subscriber pays a flat monthly fee.

21.4.2 Elements of a Modern STT

This section describes the general elements of a modern STT, including digital STTs. Of necessity, our discussion is not all-encompassing since each manufacturer does things differently, and techniques are continuously evolving as market needs change. We present this material to allow those not familiar with STTs to understand what is in one.

The Signal Path

Although variations abound, Figure 21.9 illustrates a typical block diagram of a modern addressable STT. RF from the cable enters at the top left. A directional coupler, DC1, or some other arrangement is often used to take some power in the downstream direction so as to supply signals to a data receiver if out-of-band (OOB) data is being used. Also, if an RF return path for impulse-pay-per-view service is used, it will be coupled to the cable through DC1.

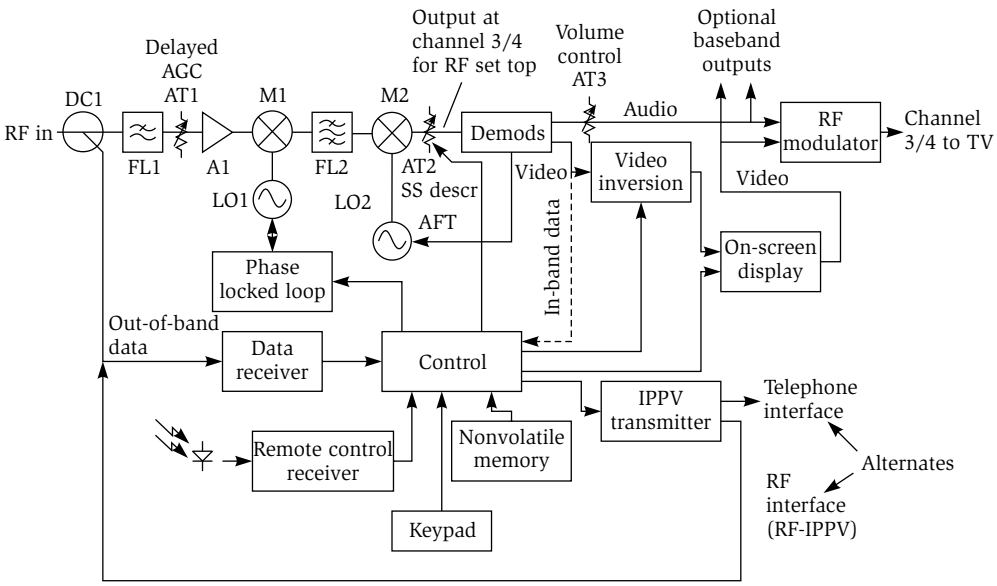


Figure 21.9 Modern addressable STT.

Following DC1 is a low-pass filter, FL1, used primarily to remove any energy above the highest frequency the set top can tune so as to reduce noise pickup and local oscillator leakage into the distribution plant. An attenuator, AT1, performs a delayed AGC operation. At lower signal levels, AT1 is at its minimum attenuation point, permitting the set top to exhibit the best noise figure it can. As signal levels increase above a certain threshold, AT1 increases in attenuation to protect the amplifier and mixer following it from being overloaded. Chapter 8 describes the delayed AGC function.

From AT1, the signal passes to a broadband preamplifier, A1, which largely sets the noise figure of the converter. The signal is up-converted in mixer M1 to the first intermediate frequency. For terminals that tune no higher than 550 MHz, a common intermediate frequency is 608–614 MHz. This would be channel 37 off-air, but no TV stations in North America are assigned to this channel, which is reserved for radio astronomy. Thus, the manufacturer can assume little likelihood of signal pickup off the air. For terminals that tune to higher frequencies, several intermediate frequencies in the neighborhood of 1,000 MHz have been used.

Up-conversion is accomplished by mixing the incoming signals with local oscillator LO1, a high-side local oscillator. Use of a high-side LO results in the signal at the first IF being spectrum inverted (see Chapter 9). The local oscillator frequency is changed to effect tuning to the desired channel. A phase locked loop is used to ensure that the local oscillator remains tuned to the proper frequency.

After filtering in FL2, the signal is down-converted either to the normal IF (NTSC picture carrier 45.75 MHz) or to channel 3 or 4 if the converter is an RF-only terminal (rare today). The second LO, LO2, may be fixed in frequency, may have limited tuning range, or may be controlled by an automatic fine-tuning (AFT) circuit, as shown here. The functions just described generally compose the tuner in the set top terminal.

The signal next encounters AT2, which does sync restoration for RF sync suppression scrambling systems, as described earlier in this chapter. This is a switched attenuator, which offers, usually, 6 or 10 dB of attenuation during active video, and is switched out during horizontal and vertical blanking intervals. Operation of the descrambler is described in Section 21.3.1.

After sync suppression restoration at IF (for the common RF sync suppression systems), the signal is supplied to video and audio demodulators. A single demodulator recovers video and audio from a digital signal. At this point, the video has had sync suppression removed, but if the video is inverted, that still must be corrected. The video inversion circuitry handles this task. As in the case of sync suppression descrambling, recovered in AT2, video inversion is controlled by the controller circuitry, as appropriate to the scrambling system being used. Digital program recovery is covered in more detail later in this chapter.

Most STTs today have an on-screen display (OSD), which is used to indicate the channel tuned, time, the program service on that channel, and other functions. In ordering pay events, the OSD is normally used. It is becoming popular to include electronic program guides in set-top terminals, and the OSD will also be used for this.

Analog audio is normally demodulated to just the composite BTSC stereo signal, as described in Chapter 8. It is possible to include a BTSC decoder. Volume control is effected at attenuator AT3, which operates on the composite BTSC stereo signal. This does not yield a truly accurate volume control mechanism since you cannot retain satisfactory stereo separation when changing the amplitude of the composite signal. The technique is used as a compromise to keep the cost low. Because it is critical to maintain the correct gain for proper recovery of BTSC stereo, most STTs include some means for normalizing the volume control setting such that the deviation out of the set top equals the deviation into the set top (unity gain). This setting is often called “best stereo,” or something similar. It is the only setting at which stereo can be recovered with optimum separation.

Volume control of the NICAM digital audio signal (see Chapter 8) is not possible without bringing it back to an uncompressed format (either digital or analog). This operation tends to be expensive, and the expense is greatly worsened if it is required to restore the NICAM signal to transmit it to the TV. Since the NICAM signal cannot be volume controlled, it is normally separated in the STT using a filter much like the one used to separate out the analog sound IF signal. After the analog video and audio are processed, and just before the RF modulator, the NICAM signal is recombined with the analog video and audio.

Terminal Control

The intelligence of the terminal is embedded in a block labeled "Control" in Figure 21.9. This block must receive addressable information from whatever means are provided by the manufacturer. If out-of-band data is transmitted, a data receiver as shown will be used to recover the data and pass it on to the intelligence of the terminal. If in-band data is used, then it must be coupled to the control section. Some terminals use data transmitted as pulses on the sound carrier using the same technique used to transmit descrambling information. Others embed control information in the VBI.

The control also interfaces with some sort of nonvolatile memory, which can retain information even when the terminal is not powered. Several technologies are in common usage. The control must receive user inputs, usually from a remote control receiver (infrared or RF), and usually from a limited function keypad on the terminal. Some sort of output must be provided to the subscriber. This output could be some combination of a numerical display on the terminal and, more likely today, an on-screen display.

Finally, the terminal must be able to communicate back to the headend. This communication is necessary to facilitate impulse-pay-per-view functions as described earlier. One way of effecting this communication involves using the subscriber's telephone line (telephone IPPV), in which case, the terminal includes a simple telephone modem. Alternatively, if two-way cable plant is available, RF-IPPV can be used, in which case, an RF transmitter sends signals upstream to the headend.

TV Display Issues

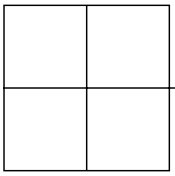
A problem with display of data on a TV set is that the TV system is optimized for displaying pictures, not small text. The compression methods used to save bandwidth, even in analog TV, don't allow for transmission of small characters, especially colored ones. The methods used to send display data from a computer to a monitor are so optimized, however. This difference in function is not always recognized, leading to frustration on the part of people who try to place too many characters on a TV screen. Generally, you should not use normal TV transmission techniques to place more than 32–40 characters on one line of text. This is contrasted to the 80 or so characters routinely used in computer work.

21.5 Summary

In this chapter, we have tried to show the most important features of scrambling systems. The most common analog scrambling technique in use in North America is sync suppression. It modifies a normal signal in ways that hide the sync. In the process, it becomes very difficult to measure signal amplitude, so this is

usually done with the scrambling turned off. Video inversion is often used in combination with sync suppression to provide a signal that is more difficult to restore. Other scrambling techniques have been used in Europe but have not enjoyed much success in North America.

An older technology, which is particularly friendly to subscribers but has limited capabilities, is trapping — either positive or negative. Interdiction has been used as an off-premises addressable technology. Most scrambling systems today are integrated into STTs, which were also described. We introduced important characteristics models. Chapter 22 will discuss digital STTs. Chapter 23 will deal with compatibility issues between cable plant and consumer equipment.



Chapter 22

Digital Set-Top Terminals and Consumer Interfaces

22.1 Digital Set-Top Terminals and Interfaces

With the widespread availability of digital video transmission on cable systems, digital set-top terminals (STTs) have taken over the marketplace. Extensive activities under the auspices of CableLabs and the SCTE, coupled with initiatives from the FCC and inputs from the consumer electronics industry, have resulted in a comprehensive set of standards for digital video transmission and reception. Many of the standards are recognized by ANSI, the American National Standards Institute. Where ANSI standard numbers were available at the time of writing, we refer to the ANSI number and also to the former standard numbers assigned by the SCTE.

A decision by the FCC to separate the STT from program denial created the need for a replaceable module to contain descrambling functions. The replaceable module is called a *POD* (point-of-deployment) security module also known as "cablecard". The idea was to allow consumers to purchase their own STTs and obtain a POD module from their cable company for reception of premium programs. In addition, TV manufacturers could choose to incorporate provision for POD modules in TV sets, eliminating the need for an STT altogether, even for premium programming. The POD module handles all vendor-specific in-band and out-of-band communications with the headend, supported by modulators and demodulators in the host device. *Host* is a generic term used to mean any device that accepts a POD module. This can include STTs, TVs, and any other device in which you might want to insert a POD module. The program input to the POD module is the demodulated 64- or 256-QAM datastream as it comes from the cable TV plant on the so-called *forward application transport* channel. (The program demodulator is located in the host.) The POD module descrambles the selected video stream from the incoming MPEG transport stream. It also handles out-of-band communications with the headend via the *out-of-band* (OOB) channel or the DOCSIS channel (advanced host). In addition, the POD module can handle additional services, such as emergency alert.

Concerns of the entertainment industry have resulted in the need for secure signal interfaces between STTs and display equipment. The ability to record signals is controlled by the copyright owner of the program. In this chapter we

outline the main issues involved in delivering digital TV to the subscriber. Of necessity, there is a tight coupling between this information and the information on digital transmission in Chapter 4 and on headend signal processing as shown in Chapters 8 and 9.

22.2 Relevant Standards

Several sets of standards contribute to the technologies described in this chapter. Most were generated under the auspices of CableLabs, in cooperation with the SCTE, and formalized into standards by the SCTE. The SCTE then submitted them to ANSI, which adopted them as U.S. standards. Additional standards have come from the Consumer Electronics Alliance (CEA), part of the Electronics Industries Alliance, and from European standards bodies.

Digital Video Subcommittee (DVS): A set of standards describing how TV signals are compressed and transmitted in cable television systems. This initiative by CableLabs has substantially completed the transition to ANSI standardization.

OpenCable: The OpenCable™ initiative was begun in 1997 with a goal of helping the cable industry deploy interactive services over cable. OpenCable provides a set of industry standards that help accomplish this goal via key objectives to define the next-generation digital consumer device, encourage supplier competition, and create a retail hardware platform.

OpenCable has two key components: a hardware specification and a software specification. The hardware specification allows a retail receiver to provide interoperability. The software specification of the OpenCable project, called the *OpenCable Applications Platform* (OCAP), is designed to solve the problem of proprietary operating system software by creating a common software platform upon which interactive services may be deployed.

VOD metadata: This is a newer initiative. It is a project to investigate the distribution of content assets (for example, movies) from multiple content providers and sent over diverse networks to cable operators. Metadata is descriptive data associated with a content asset package. It may vary in depth from merely identifying the content package title, to information for populating an EPG, to providing a complete index of different scenes in a movie or providing business rules detailing how the content package may be displayed, copied, or sold.

PacketCable: PacketCable is a CableLabs-led initiative aimed at developing interoperable interface specifications for delivering advanced, real-time multimedia services over two-way cable plant. Built on top of the cable modem infrastructure, PacketCable networks use Internet Protocol (IP) technology to enable a wide range of multimedia services, such as IP telephony, multi-

media conferencing, interactive gaming, and general multimedia applications. Much of PacketCable is covered in Chapter 6. Another early and continuing initiative, DOCSIS, is also covered in Chapter 5.

22.3 Cable TV Digital In-Home Processing

We could have described Figure 22.1 as a representative high-end digital STT, but it doesn't necessarily represent a real device.[1] Rather, it could be looked at as a superset of most real digital STTs or any device that accommodates POD modules. We'll refer to this as a *host*, an all-encompassing term used to describe any device that accommodates a POD module. You may not find all of the features shown on every host.

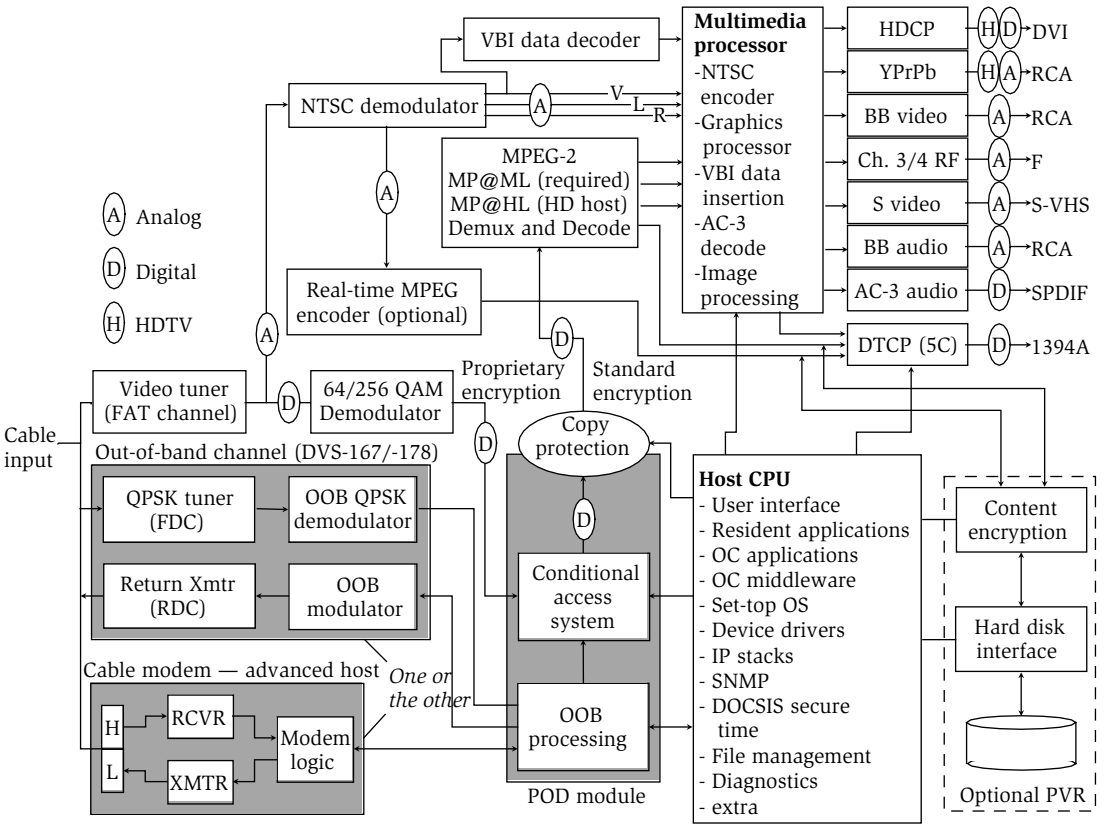


Figure 22.1 Representative digital host.

Many STTs are built with the POD module functions internal to the STT, so a separate POD module is not necessary. However, the intent of the FCC rules is to promote the availability of retail STTs that will use a POD module for customization of scrambling. The idea is to allow the consumer to purchase an STT, TV, or whatever with the features he or she wants and to get only the POD module from the cable operator. Also, it is possible to build TV sets, VCRs, and other consumer devices with POD module sockets, to allow them to operate as standalone devices, but to use the POD module to descramble digital signals. Note that the POD module is required only for scrambled programming; programs sent in the clear don't need a POD module to be installed.

22.3.1 Block Diagram Description

Video signals are tuned in the video tuner, which can be the same double conversion tuner used in analog STTs. Figure 21.9 illustrates a tuner in the blocks from the RF in through the second mixer M2. Dual conversion is used to achieve flat response over a very wide range of frequencies while not having to reject an in-band image. In Figure 22.1, the ovals with D or A in them indicate routes taken by digital or analog signals, respectively. An oval with an H indicates a high-definition signal path, which may also carry standard-definition signals.

Analog signals are routed to an NTSC demodulator of conventional design. A VBI decoder recovers VBI data to be inserted into the video signal in the multimedia processor. VBI data includes closed captioning text sent in the VBI of an analog signal, usually on line 21.

The multimedia processor accepts video in several forms as well as accepting VBI data from the VBI data decoder. Conceptually, the signals coming from the MPEG-2 decoder may or may not be in NTSC format (depending on the decoder used). Normally the multimedia processor will convert any format to RGB for insertion of graphics. The multimedia processor converts the signal to NTSC for conventional display and to whatever other formats are required, as described later. The graphics processor is the onscreen display of Figure 21.9, which adds VBI text data when called for and also displays text and graphics generated in the host. An example of host-generated graphics would be the grid of an electronic program guide.

The output from the multimedia processor may take a number of forms, depending on the capabilities intended for the host. We'll discuss the interfaces in more detail later, but can name them briefly here.

DVI is a digital interface used to transmit non-compressed HDTV pictures to a display. HDCP describes the copy protection specified for use with DVI. The YPrPb display is an analog component interface that is popular in early HDTV displays. It couples luminance (Y) and two color difference signals (Pr and Pb) to the display device.

Baseband, or BB, video is conventional analog NTSC (or PAL or SECAM) baseband video, as described in Appendix B and in Chapter 2. Ch. 3/4 RF output

is again the conventional RF interface used to couple signals to TVs that don't have any other interfaces available. Generally, the quality of the RF interface will be the lowest quality of any interfaces, though in many cases you will be hard pressed to see the difference between the RF interface and the baseband video interface. S-VHS is an analog interface that is the same as baseband video except that the luminance and chroma are transferred on separate wires. This has some distinct advantages in allowing for wider luminance bandwidth and in elimination of interference between the luminance and chroma information.

Baseband audio is conventional one- or two-channel analog audio for use by stereo systems. SPDIF is a digital audio interface originated by Sony and Philips (the Sony–Philips digital interface). It is a common digital interface for digital audio using the AC-3 format specified for digital transmission in North America.

Note that there is one digital interface that does not originate in the multimedia processor: the 1394 interface specified for standard-definition digital interfaces. DTCP is the copy protection system specified to be used with 1394 display interfaces. The reason it isn't shown originating from the multimedia processor is that it is able to transfer the MPEG-compressed video from the incoming video stream, plus information telling the display how to overlay any characters or graphics intended for display (the overlay information is provided from the multimedia processor). Thus, when using 1394, the display device, rather than the source device, is required to form the composite signal by overlaying graphics and/or characters on the picture. The multimedia processor does not have to assume this responsibility.

Returning now to a brief tour of Figure 22.1, an optional real-time MPEG encoder is shown. It is used if the host wants to be able to convert NTSC signals into digital signals for transmission to the display. This could be useful to allow consistent 1394 interfaces to the display. Another reason for including a MPEG encoder would be to accommodate a hard disk drive for storing video. The host would need such an encoder to accommodate analog signals. The storage function has come to be known as personal video recorder (PVR). Note the content encryption at the input to the hard disk interface. This is to ensure that content captured on the hard disk can be accessed only by the user who was authorized to capture the video and to limit its period of use if desired.

The POD module controls decrypting of digital video-encrypted signals. All of the program denial functions that allow the operator to sell premium services are embedded in the POD module, so the host without a POD module is useless for receiving anything except basic service.

Two options for out-of-band (OOB) communications are shown. The out-of-band channel is a QPSK tuner and demodulator plus a QPSK modulator and transmitter conforming to one of two OOB communications protocols recognized by the SCTE. The two are described later. The second OOB communications channel, anticipated for so-called advanced hosts, is a DOCSIS modem, as described in Chapter 5. The idea is to build in a DOCSIS modem, which handles host communications needs. The host could have a data interface to allow the subscriber to use this DOCSIS modem for all data communication needs.

Finally, the host CPU is the computer that controls all the functions of the host. These functions include user interfaces, such as the remote control and on-host controls, as well as any onscreen or on-host displays. The CPU will also house some resident applications that ship with the host. OCAP middleware running on the host CPU allows external applications to be written and loaded to the host. Other functions include external communications, the host role in network management, management of files, diagnostics, and whatever other functions the host manufacturer wants to include.

22.3.2 Cable Input

The cable input is standard cable TV signals, including return signals exiting the host. Input signals range from 54 to about 864 MHz, possibly increasing in the future. (The low frequency of 54 MHz is a North American standard. Other localities often use higher minimum downstream frequencies, allowing a wider upstream bandwidth.) The controlling document is SCTE 40 2001 (formerly DVS 313).[2] This document further references EIA-23 and EIA/CEA-542-B, specifications developed by the former NCTA/EIA Joint Engineering Committee. EIA-23 defined certain analog interfaces between cable systems and consumer electronics equipment. EIA/CEA-542-B defines channelization. The channel boundaries stated in this specification are reproduced as Appendix A. SCTE 40 2001 defines several types of signals that can be present on the cable plant:

NTSC: These are conventional analog TV channels on the cable.

FAT: This unfortunate acronym stands for *forward application transport* channels, which carry digital information via MPEG-2 transport streams. These are the digital video channels. Modulation is either 64- or 256-QAM.

NTSC and FAT channels may occupy the spectrum from 54 to 864 MHz and are channelized according to the table in Appendix A. (If history is any indication, the maximum frequency will increase in the future.) Signal level is measured at the end of the drop, except as noted. The maximum signal level is +15 dBmV for analog and FAT channels. For analog signals, the minimum signal level is 0 dBmV (governed by FCC rules) at the first consumer-owned device. Digital signal levels are carried at lower levels because they don't need as high a carrier-to-noise ratio (C/N) as do analog signals. 64-QAM modulated signals are carried nominally −10 dB with respect to the level of a hypothetical analog carrier on the same frequency. 256-QAM signals are carried nominally −5 dB from the hypothetical analog level. The specification allows a wider range of signal levels at the consumer device for digital signals, −15 dBmV to +15 dBmV for 64-QAM and −12 dBmV to +15 dBmV for 256-QAM.

C/N is an important quantity used to judge the quality of a signal delivered to subscribers. Today, C/N is sometimes replaced by the *carrier-to-composite-noise* ratio (CCN) or *carrier-to-noise-plus-interference* ratio (C/(N + I)). These

terms recognize that nonlinear distortions and leakage, which introduced essentially CW interference in analog systems, introduce interference that has the characteristics of noise in digital systems. C/N for an analog NTSC signal should be better than 43 dB (controlled by FCC rules), though good engineering practice dictates higher numbers. For 64-QAM the C/(N + I) should be no less than 27 dB, and for 256-QAM it should be no less than 33 dB. Chapter 4 explains why.

FDC: Out-of-band forward data channels carry control information to STTs and other devices that need it. This is also known as the downstream out-of-band (OOB) channel for addressable STTs. The information transported includes STT channel and program information (collectively known as *system information*, SI) and STT queries. The main information transmitted are *entitlement management messages* (EMMs), used to tell the STT or POD module what services it is entitled to receive. FDCs are usually modulated with QPSK (4-QAM) and set to −8 dB with respect to a hypothetical analog picture carrier at that frequency. FDCs are carried between 70 and 130 MHz.

RDC: Out-of-band reverse data channels carry return information from the STT and other devices to the headend. These channels may be carried anywhere between 5 and 42 MHz, depending on the cable system. Though reverse channel output levels may be up to +60 dBmV coming out of a device, the level allowed to appear at the input to another device is limited to +42 dBmV.

22.4 Out-of-Band Channels

The out-of-band FDC and RDC channels may take any of several forms. For advanced hosts, a DOCSIS modem may be included in the host to handle OOB communications. If a DOCSIS modem is included, the host may have an external interface for the DOCSIS modem to allow it to serve data needs for the entire home, not just for the host. The legacy OOB data channel shown in Figure 22.1 is based on one of two specifications: DVS-178 (SCTE 55-1 2002) or DVS-167 (SCTE 55-2 2002). Table 22.1 illustrates key features of the two standards: both the OOB FDC and the OOB RDC. SCTE 55-1 was originally developed by General Instrument; SCTE 55-2 was adapted by Scientific Atlanta from the DAVIC standard. In addition to two-way communications, it is important to provide for a one-way communications mode so that if the return channel fails for any reason, basic broadcast functionality, such as EPG, may be maintained.

22.4.1 Downstream OOB Channels

In the downstream direction, SCTE 55-1 operates at a wire rate (the actual data rate on the coax) of 2.048 Mb/s, with an information data rate of 2.005 Mb/s. SCTE 55-2 has two downstream grades, or modes, of which grade A is the only

Table 22.1 Comparison of out-of-band standards

	SCTE 55-1	*SCTE 55-2*
Formerly	DVS-178—mode A	DVS-167—mode B
Based on	MPEG, ATM	DAVIC
Downstream		
Data rate	2.048 Mb/s	1.544 Mb/s (grade A, required) 3.088 Mb/s (grade B, optional)
Nominal information rate	2.005 Mb/s	1.28 Mb/s
Frequency	70–130 MHz	70–130 MHz
Channel spacing	1.8 MHz	1 MHz (grade A) 2 MHz (grade B)
Modulation	D-QPSK, raised cosine $\alpha = 0.5$ (filter at receiver)	D-QPSK, root raised cosine $\alpha = 0.3$ (filter split)
Error correction	Randomization, R-S, interleaving	R-S, interleaving, randomization
Received carrier power	−10 to +5 dBmV	−18 to +15 dBmV
Segmentation and reassembly	Based on MPEG-2	ATM AAL5
Addressing	Broadcast Singlecast_unit Singlecast_network Multicast 40, −16, 24	Broadcast Singlecast
Upstream		
Data rate	256 kb/s	Grade A (optional): 256 kb/s Grade B (required): 1.544 Mb/s Grade C (optional): 3.088 Mb/s
Modulation and filtering	D-QPSK, $\alpha = 0.5$	D-QPSK, $\alpha = 0.3$
Access scheme	Polling, Aloha	Slotted TDMA synchronized to a common clock, reserved and contention slots
Packetization	ATM cell preceded by 28-bit unique word and 1-byte packet sequence, followed by 8 bytes R-S parity	ATM cell preceded by 32-bit unique word, followed by 6 bytes R-S parity and 1 byte guard band

Frequency	8.096–40.160 MHz	8–26.5 MHz
Channel spacing	192 kHz	Grade A: 200 kHz Grade B: 1 MHz Grade C: 2 MHz
RF output power	+24 to +60 dBmV	+25 to +53 dBmV
Received level at headend	$+3 \pm 10$ dBmV	Four target settings, −5, +3, +11, +19 dBmV, all ±3 dB
Required C/N at receiver	20 dB @ PER $< 10^{-7}$ after FEC	20 dB @ 10^{-6} packet loss rate after error correction
Acknowledgment protocol	STT waits for acknowledgment from headend or enters backoff algorithm	Positive acknowledgment sent for every successfully received ATM cell
Maximum upstream channels per downstream channel	6	8

one implemented as of this writing. It features a downstream wire rate of 1.544 Mb/s with a data throughput rate of 1.28 Mb/s. Both standards include the same downstream frequency range of SCTE 40, though in addition SCTE 55-1 shows a default frequency of 75.25 MHz.

Note a difference in the downstream filtering in the two standards. In Chapter 4 we showed that most digital transmission systems employ raised cosine bandpass filtering and that most split the filter evenly between the transmitter and the receiver. This leads to the description of a root raised cosine filter (the "root" terminology arises from the fact that the filter response at each end is the square root of the overall filter, so when the transmit and receive filters are cascaded, a raised cosine filter results). This is the technique used in SCTE 55-2, but SCTE 55-1 places all the filtering at the receiver. This results in a wider occupied spectrum on the cable but improves the C/N characteristics of the receiver.

The numbers in Table 22.1 were taken from the SCTE 55-1 and -2 specifications. There is a difference between them and the downstream received signal level specification of SCTE 40, but the difference is not of practical concern if you follow the recommended practice of placing the data carrier −8 dB from a hypothetical analog picture carrier at the same frequency. Note also that the downstream addressing for both standards allows addressing either individual hosts or all hosts at once.

22.4.2 Upstream SCTE 55-1 Channel

In the upstream direction, SCTE 55-1 uses a single data rate of 256 kb/s. The difference in upstream access schemes is of interest. SCTE 55-1 has a polling channel. A polling protocol may be thought of as a "speak only when spoken to"

protocol. Each host is polled and responds when polled and only when polled. This polling channel (which can be a separate return frequency) is used to allow the host to report its health and to report any pay-per-view programs that have been watched since the previous poll. Other non-time-critical data may be transmitted on the polling channel. Polling takes place on an infrequent basis — each Host may be polled perhaps once per day.

SCTE 55-1 also specifies one or more Aloha channels. These may be used by different applications running on the host. For example, a video-on-demand (VOD) system would like to use the host to request programs at any time and would also like to allow the viewer to command the program to pause and resume, rewind and fast-forward (so-called VCR functions). Aloha plays a central role in the vast majority of Ethernet networks in use today. The Aloha Protocol was developed at the University of Hawaii by Norman Abramson and colleagues in the late 1960s as a way to communicate by radio between the various islands and interisland shipping interests. The radio channels used in Hawaii had a lot in common with the return channel of cable systems today. The Aloha Protocol allowed any station to transmit when it needed to transmit, rather than having to wait until it was "invited" to transmit, using polling techniques. There was no guarantee that one station would be able to hear any other station that was transmitting at the same time. In the original application, this was because of the vagaries of radio paths. In cable plant, there is no way to route a signal from one host to another, so when one host is transmitting, no other hosts will know it.

In a pure Aloha system such as in SCTE 55-1, when a host has something to say, it transmits and then waits for an acknowledgment from the headend. If no one else is using the channel, the transmission gets through, the headend acknowledges it, and the host goes on about its business. If an acknowledgment is not received within a designated time, the host assumes its message didn't get through because another transmission at least partially overlapped with its transmission. This means there is at least one other host trying to communicate. So the host waits ("backs off") a random time and tries again. Chances are that the other host who was transmitting at the same time has backed off for a different random time and that on the next try both will get through and will be acknowledged.

Pure Aloha systems work well when there are a lot of potential "talkers," each needing the channel a small percentage of the time but needing it quickly when it is needed. A VOD application is a good example. Normally, VOD systems allow the subscriber to have VCR-like control, in that they can pause and resume the program, rewind, or fast-forward. A lot of subscribers can be watching VOD and sharing a return channel to send those commands to the headend. Rarely does any one subscriber want to use any of those VCR-like controls; but when he does, he needs a response fast from the headend.* Polling wouldn't work, because even

---

* There are some instances involving networked PVRs in which a number of people will want to simultaneously access the VCR-like controls, stressing Aloha.

fast polling of many hosts could take minutes if not hours. A VOD viewer would not wait that long to get a response to, say, a pause command. But since there are few commands issued by each subscriber, many can use the channel and not experience significant interference. A pure Aloha system can be utilized to about 18% before delays due to collisions become unacceptable.[3]

Figure 22.2 illustrates a typical STT communications channel based on SCTE 55-1 (formerly DVS-178). Some cable TV systems are intended to be locally controlled in terms of STT authorization, and these systems use a control computer. Other systems using SCTE 55-1 are designed to be controlled from a central point via satellite. These systems don't use a control computer but, rather, have a master integrated receiver-transcoder (IRT—a digital satellite receiver) that supplies control information. Either the master IRT or the control computer supply signals to an OOB modulator, which sends the downstream QPSK signal as described in Table 22.1. This signal is combined with all other downstream signals and supplied to the distribution network, commonly via a downstream optical transmitter.

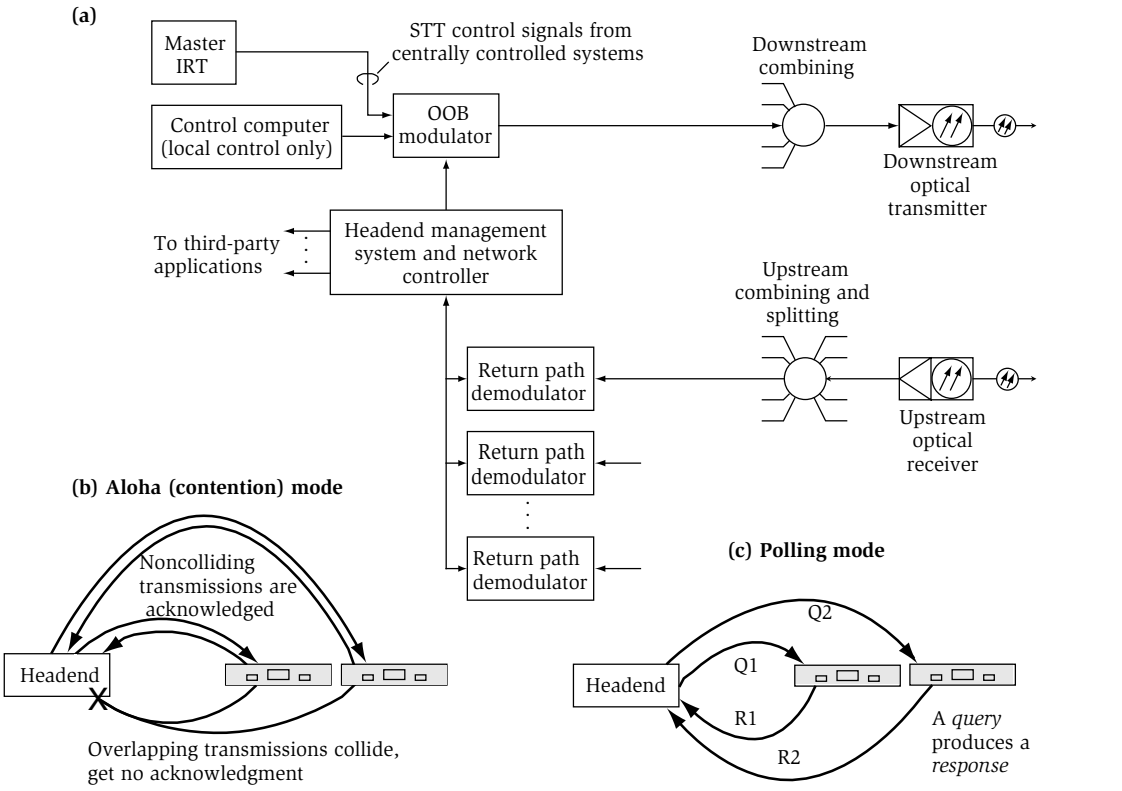


Figure 22.2 SCTE 55-1 communications system.

In the upstream direction it is common (though not universal) to make provision for several upstream return path demodulators for each downstream modulator. Upstream signals, either polling or Aloha Protocol, come to the headend, usually on an upstream optical receiver. Since there may be multiple combinations of upstream signals for different services, as described in Chapter 9, there will be some combining and splitting loss in the signal path before the upstream signals reach the assigned return path demodulator. The demodulators used may be assigned to different groups of STTs (by the way the upstream signal management is done), or one receiver may be assigned as the polling receiver and the others assigned as Aloha Protocol receivers, each supporting a different service (such as VOD or game playing). The receivers may be assigned by virtue of being tuned to different upstream frequencies. For example, the polling receiver might be assigned to 8.096 MHz (the lowest channel) and the VOD receiver (Aloha Protocol) to 8.288 MHz, the next channel up.

Each receiver demodulates the QPSK upstream signals and supplies it to the headend management system and the network controller. The headend management system receives the information on the polling channel and handles interfaces with the billing system and other systems needing access to the information. The network controller provides interface to third-party applications, such as a VOD system.

22.4.3 Upstream SCTE 55-2 Channel

The system used in SCTE 55-2 (formerly DVS-167) for return transmission is significantly different from the SCTE 55-1 system. It is called a time division multiple access (TDMA) system. SCTE 55-2 allows for three different upstream data rates, though only one is implemented as of this writing. Figure 22.3 shows the overall system. Systems using SCTE 55-2 were generally designed to operate with local host control. A digital network control system (DNCS) provides physical interface to applications needing access to the STTs. The control transmitter is the QPSK transmitter that generates the downstream signals shown in Table 22.1. As with SCTE 55-1 systems, several receivers may be used to improve the upstream throughput.

The upstream receivers demodulate the incoming signals and pass them to the control transmitter, which in turn passes signals to the DNCS. The DNCS manages some of the communications with hosts itself and relays other communications to and from other computers running various host tasks.

Upstream transmission is synchronized from a common clock at the headend. Downstream transmissions are divided into repeating signaling link extended superframes (SL-ESFs). Each SL-ESF consists of 24 frames, each carrying 193 bits (1 synchronizing bit and 192 data bits). Timing for the upstream begins with receipt of *SL-ESF START* from the headend. Upstream time slots are defined from this start of the downstream SL-ESF. Normally each Host that needs to talk reserves one or more of the upstream time slots in which to talk. Time slots may be assigned for one-time use or may be reserved for a longer duration

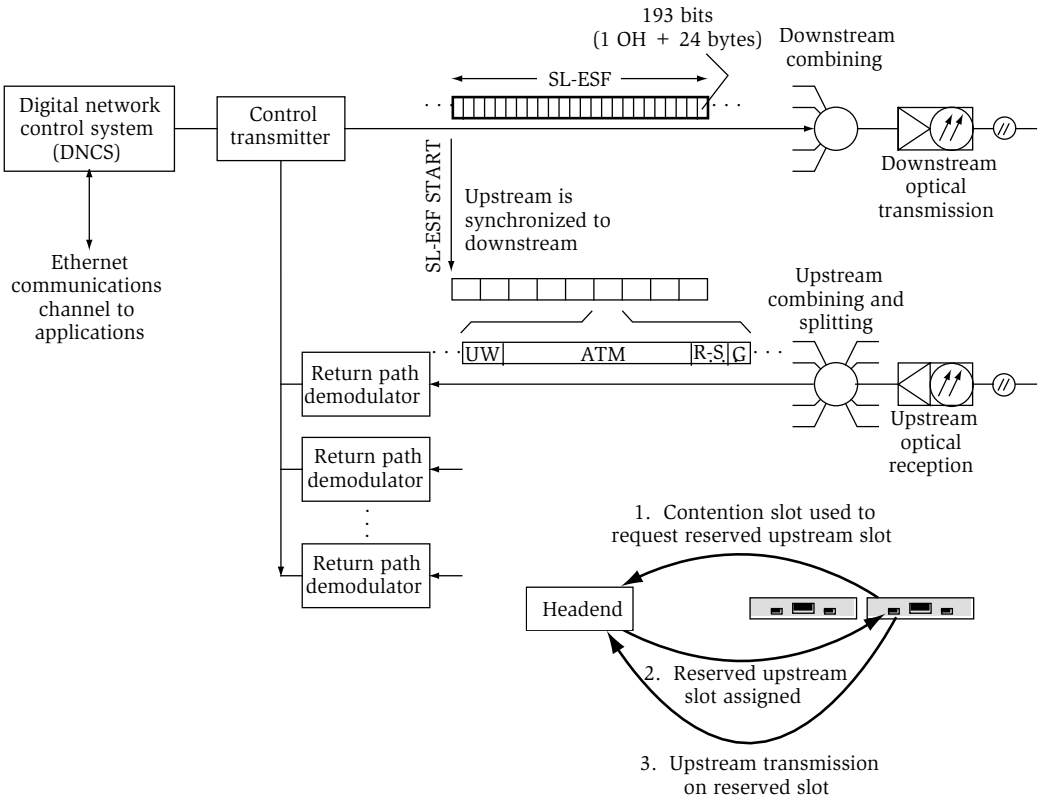


Figure 22.3 SCTE 55-2 communications system.

if an application needs it. An application may also reserve adjacent slots. For upstream transmission at 1.544 Mb/s, the most common speed, there are nine upstream time slots for each SL-ESF. These time slots consist of a unique word (UW) that serves as a run-in to synchronize the receiver clock. Following this is a 53-byte ATM cell. This is followed by a Reed–Soloman error correction word and finally by a guard time, which ensures that small timing errors don't cause transmission overlap from different hosts.

Propagation time between the headend and the host is important, since it is necessary to time all transmissions to arrive at the headend at the proper time. The propagation delay difference between a host that is close to the headend and one that is distant is significant, so a ranging function must be applied to each host when it is first added to the system. The headend measures the time delay between the "expected" response time from the host and the actual response time, and it tells the host how much to advance its timing so that the signal

arrives at the headend at the correct time. Other systems that depend on tight timing also have this ranging function.

In order to allow a host to gain access to a time slot, there must be provision for a host to inform the headend that it needs a time slot. Also, when a new host is added to the system, it is necessary to allow it to inform the headend of its presence. In order to meet these needs, a return time slot is occasionally set aside for contention. This time slot will work very similarly to the Aloha system described earlier, except that rather than the host's having the ability to begin "talking" at any arbitrary time, it must wait for a contention time slot, which is defined by the headend. This modification of the pure Aloha Protocol is called *slotted Aloha*. There is a time slot set aside periodically in which any host needing access to a reserved time slot so informs the headend in the contention time slot. The headend acknowledges the request and assigns an available time slot for that host. Whereas a pure Aloha system becomes unusable at about 18% utilization, slotted Aloha can operate with a utilization of up to about 37%, because there are no partial time overlaps.

The inset drawing at the lower right of Figure 22.3 illustrates the sequence of events that a host must go through in order to deliver an upstream message. Step 1 is to wait until a contention slot comes along. The host transmits its request for an upstream time slot to the control transmitter, which manages the assignment of time slots. In step 2, the control transmitter assigns a time slot to the host. In step 3, the host transmits its data. The upstream time slot is normally assigned in the second SL-ESF following the request. Some short messages may be sent in the contention slot rather than going through the process of getting an assigned time slot.

22.4.4 Advanced Hosts

As shown in Figure 22.1, a DOCSIS communications channel may be substituted for the SCTE 55 communications channel. This so-called DOCSIS set-top gateway, DSG, uses Ethernet tunneling to transmit data to the host, and so it doesn't need the QPSK channel described earlier. See Chapter 5 for a discussion of DOCSIS modems. If this is done, it is also possible that the host will include an interface to computers, so the modem may be used both for host communications needs and for computer communications. Possible interfaces include Ethernet, USB, and IEEE 802.11, among others.

22.5 Output Interfaces and Copy Protection

A number of output interfaces are shown for the model host of Figure 22.1. We shall first cover the various analog interfaces that are popular on early-generation digital STTs; then we shall cover the digital interfaces.

22.5.1 Analog Interfaces

BB Video

This is the common baseband analog video interface described in Chapter 2 and Appendix B. Sound is not included. The significant frequencies extend from nearly 0 to 4.2 MHz for NTSC video and from nearly 0 to 5 MHz for PAL video. Composite video (including sync, as described in Appendix B) is transmitted at a level of 1 volt p-p, with a source and load impedance of 75 ohms. Normally, the common "RCA" connectors are used, though occasionally other connectors, such as BNC, are employed. Impedance matching is important if the connecting cable is long, to prevent ringing or ghosting. The quality of the interface is better than that of the Ch. 3/4 interface described next but not as good as the other analog interfaces described, because the chrominance information is combined with the luminance, and the two must be separated before they can be processed. The filters used to separate the two are imperfect; even if perfect filters were available, there would be some inevitable interference.

Ch. 3/4 RF

This is baseband video modulated onto an RF carrier, identical to signals transmitted on cable systems and over the air. The one exception is the lower sideband, which is normally removed by the vestigial sideband filter (described in Section 8.3.7). Vestigial sideband filtering is done to allow adjacent 6-MHz channels to be used. Often the vestigial sideband filter is omitted in consumer products. This has no effect on the quality of the signal received, but it does preclude placing another signal in the lower adjacent channel. Also due to cost, the sound carrier frequency may not be as accurate, and the picture/sound carrier amplitude ratio may vary. Again, these differences should not pose a problem in consumer applications.

Although the quality of the Ch. 3/4 video interface can be quite good, of all the interfaces available, this one would be considered the lowest in quality, since it starts with composite video and modulates it onto a carrier, requiring demodulation in the receiver. The extra modulation and demodulation steps cannot improve the quality of the signal.

In North America it is common to use channel 3, except in areas having an off-air channel 3, in which case channel 4 is used. Channel 2 is not commonly used because it is the second harmonic of the old citizen's band radio service (CB), and harmonics of CB services have been known to impair reception.

The signal level coming out of the STT should be above 0 dBmV and less than approximately 15 dBmV. Sound is modulated onto a subcarrier at 4.5 MHz (NTSC), 5.5 MHz (PAL-B/G), or 6 MHz (PAL-I). F connectors are used almost universally in consumer equipment. It is important to ensure that the source and load impedances are 75 ohms and that 75-ohm cable is used, regardless of the length of the cable.

S-VHS

This standard originated in the mid 1980s as S(uper)-VHS, an improved version of VHS consumer videotape. The recording standard did not become popular for consumer use, though it has been used for semiprofessional purposes. Compared with normal VHS recording, there were some differences in the way the recording was made on the tape. The interface is now sometimes called S-video.

Luminance and chrominance (color) signals are transmitted to the display on different wires. The luminance information is transmitted as a baseband 1-V p-p signal including sync, just as in baseband video. The color subcarrier (3.58 MHz in NTSC and 4.43 MHz in PAL) is not superimposed on the luminance as it is in normal baseband video. Rather, it is transmitted on a different wire. The advantage is that no filtering is needed to separate the luminance and chrominance information. This permits wider bandwidth of the color signal, improving color resolution. Many TVs today have S-VHS inputs, and excellent quality can be achieved using this interface.

Besides allowing wider chrominance and luminance bandwidths, S-VHS eliminates cross-color effects that arise when the chrominance and luminance signals interfere with each other. Cross-color effects may be seen, for example, when fine diagonal stripes, such as in a striped jacket, are on the screen. You have seen spurious colors surrounding such images. Spurious color occurs when luminance signals with a lot of high-frequency content (as are generated by diagonal stripes) mix with the color spectrum, confusing the color demodulator. Since properly produced digital video is derived from a signal that has never existed in composite analog video format, use of S-VHS connection eliminates cross-color effects.

S-VHS uses a four-pin miniature connector. As shown in Figure 22.4, the luminance (Y) and chrominance (C) each have their own ground. The luminance signal is identical in every respect with the baseband luminance signal, including sync. The chrominance signal is a modulated 3.58-MHz (NTSC) or

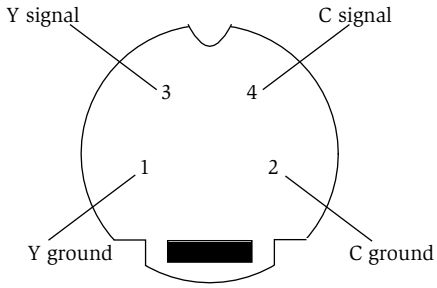


Figure 22.4 S-VHS pin function.

4.43-MHz (PAL) signal identical with the color subcarrier of the composite baseband signal. If you combine the Y and C signals, you will have a proper composite baseband signal. Again, the reason for the S-VHS connector is to keep the two signals separate so that they will not interact. A second improvement is that separating the chrominance from the luminance allows a wider chrominance bandwidth, resulting in sharper color edges.

Audio is not carried on the connector and is normally carried as left and right audio on a pair of cables terminated in RCA jacks.

YPrPb

This is an analog component interface using three cables to transfer the luminance (Y) and red and blue difference signals (Pr and Pb). It is used for standard definition video and may optionally be provided for HDTV display interface. Sync is carried on the luminance channel. Analog copy protection may be included, as described later. Three separate coaxial cables are used, often bound into a single cable. For consumer applications, RCA connectors are used, whereas BNC connectors are frequently used in professional applications. As with other video signals, 75-ohm cable should be used, though non-impedance-controlled cable is sometimes used in consumer applications.

As shown in Chapter 2, a color television signal is composed of red, green, and blue components. For transmission, these components are converted to a wideband luminance (Y) signal and two narrowerband color difference signals. When the signals are transmitted in analog form, the color signals are called the Pr and Pb signals, for the luminance minus red and luminance minus blue components. The same components, when in digital form, are called YCrCb.

The components are defined by CCIR Recommendation 601 for modern systems as follows:

$$
\begin{aligned}
Y &= 0.299R + 0.587G + 0.114B \\
P_B &= \frac{0.5}{1 - 0.299}(B - Y) \\
&= -0.169R - 0.331G + 0.5B \\
P_R &= \frac{0.5}{1 - 0.114}(R - Y) \\
&= 0.5R - 0.419G - 0.081B
\end{aligned}
$$

The physical interpretation is as follows. Assume the red, green, and blue (RGB) outputs from the camera range from 0 for black to 1 (the units can be volts or some arbitrary unit) for maximum of each primary color (the camera must be adjusted such that these maximums correspond to the "brightest" color to be reproduced). Then at each pixel (picture element, or the smallest resolvable element in the scene), R, G, and B will each have a value. The luminance (Y)

signal is computed from the first line of the equation, which approximates the response of the human visual system to each color. In order to make up the two color difference signals, the difference between Y and the R and B channels is taken, with the scaling factors shown. Notice that the color difference signals can take on negative values, even though there are no real negative colors. The two color difference signals usually have one-half the bandwidth of the luminance signal (less in analog transmission systems).[4]

The foregoing equations are usually written in matrix form,

$$\begin{bmatrix} Y \\ P_B \\ P_R \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.169 & -0.331 & 0.500 \\ 0.500 & -0.419 & -0.081 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

Analog Copy Protection

An analog copy protection system was developed a few years ago to prevent consumers from recording premium programs on VCRs. It is used to prevent recording on analog video outputs from the host as well as on DVD players. The system takes advantage of the fact that VCRs use AGC on the video signal to set the peak-to-peak amplitude of the video signal before recording. Special signals are inserted in the vertical blanking interval that cause the AGC to set up to the incorrect level part of the time.

BB Audio

This interface is conventional analog transmission of baseband audio. In stereo systems, left and right channel information is transferred. For home theater use, 5.1 channel audio can be transmitted on six cables. The channels are left, center, and right front channels, left and right rear channels, and a low-frequency enhancement (LFE) channel. The LFE channel is limited to approximately 20–100 Hz and hence is considered "0.1 channel." Used to provide "feel" to movies in home theaters, it is an optional feature of digital TV. (There is a newer practice that uses a one or two rear channel for 6.1- or 7.1-channel audio, but this is not officially supported in most interfaces as of this writing.)

Common interface levels are about 1 V rms at full modulation. The terminating impedance is usually 2 kilohms or higher, and the source impedance is much lower. Because the frequencies of audio are so low, there is no need to maintain impedance match in consumer systems, and maintaining a much higher load impedance than the source impedance enables more than one load to be placed in parallel is desired. Compare this practice with professional equipment, where a load impedance of 600 ohms is used. The source impedance is often, but not always, either 150 or 600 ohms. Though the impedance of the cable is not

critical, it is important to use a high-quality shielded cable to minimize hum pickup and crosstalk.

Consumer equipment almost always uses unbalanced transmission, whereas professional equipment almost always uses balanced transmission, drastically improving the immunity to hum and spurious signal pickup. The issues related to balanced and unbalanced transmission are covered in Section 8.9.2. If long cables are used, there is some possibility of high frequency roll-off due to cable capacitance. This is normally not a serious problem in small home entertainment clusters, but if you find yourself running audio cables between rooms or connecting them with Y adapters, you should pay attention to the total cable capacitance. With a source impedance of 2 kilohms and a high load impedance, a total cable capacitance of some 800 pf could marginally impair frequency response (we are being conservative). Typical audio cables seem to exhibit capacitance of about 30 pf per foot, so as the total cable distance approaches 25 feet per channel, it could introduce some very minor response problems if the source has a 2 kilohm output impedance.

22.5.2 Digital Interfaces

DVI

This is the interface specified for high-definition digital displays. Uncompressed RGB (red-green-blue components, as opposed to luminance and color difference) digital video is transmitted, so the MPEG decoder is in the host. DVI has its roots in computer monitors. Manufacturers recognized that with high-resolution monitors the analog interfaces that had been used for many generations of product were now limiting performance. Furthermore, with LCD displays, the signal to be displayed was converted to analog in the computer and then sent to the display, where it was converted back to digital. DVI was created by the Digital Display Working Group[5] (DDWG) to convert analog signals into digital signals to accommodate both analog and digital monitors. The standard specifies a single plug and connector that encompass both the new digital and legacy analog interfaces as well as a digital-only plug connector. However, for television applications, the analog channel is not used. DVI handles bandwidths in excess of 160 Mb/s per channel and thus supports UXGA and HDTV with a single set of links. Higher resolutions can be supported with a dual set of links (not used for HDTV applications).[6,7]

The interface uses a physical transmission protocol known as transition minimized differential signaling (TMDS). In concept it is similar to 8B/10B encoding used with gigabit Ethernet and described in Chapter 19. Two additional bits are added to each word for the purpose of minimizing the number of transitions (in this case) while removing any DC component of the signal.

Figure 22.5 summarizes information concerning the physical interface used in DVI. Figure 22.5(a) shows the transmission of signals between a host such as

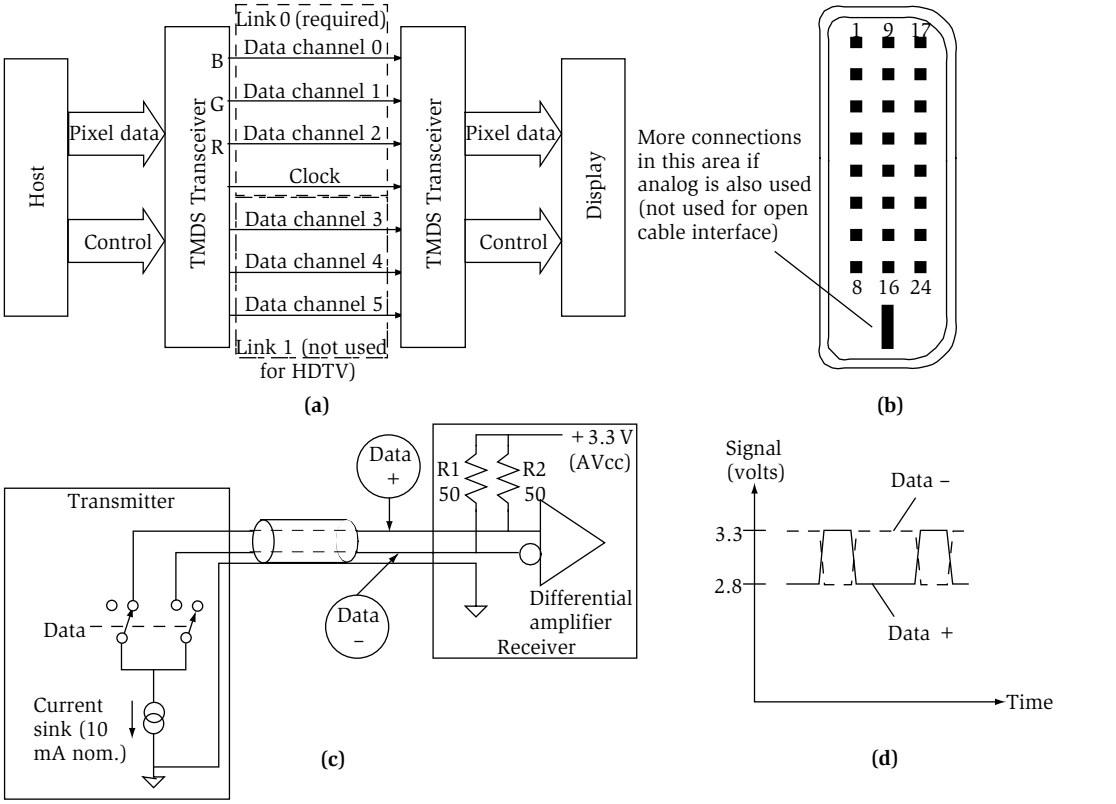


Figure 22.5 The DVI. (a) Communications channels. (b) DVI plug. (c) Equivalent electrical circuit. (d) Waveform.

an STT on the left and a display on the right. Pixel and control data are transferred to a TMDS transceiver and across a data channel (a DVI cable) to a TMDS transceiver on the display, which converts the signals back to pixel and control data for the display. Two complete digital links are included in the DVI specification, called link 0 and link 1. Each link consists of three data channels, for blue, green, and red data. A common clock line is used for all channels. The reason for two links is to double the effective data transfer speed. Only link 0 is needed for HDTV applications. Data channel 0 carries blue information, data channel 1 carries green, and data channel 2 carries red information. Horizontal and vertical sync information is carried in the blue channel. (Note that this is a different practice from that normally used in analog component video, where the green channel carries sync. Green is the logical choice in analog practice because the picture is made up predominantly of green. In digital, there is no particular reason for the sync to appear on one channel over any other.)

DVI can operate at data rates from 25 Mb/s to 165 Mb/s, depending on how much data is required to be transferred. At the same frame rate, there is up to 5.3 times as much information that must be transferred for HDTV as for SDTV, depending on the display format being used.

Figure 22.5(b) illustrates the connector used for DVI. Though not specified for HDTV use, there is an option for conventional analog signals in accordance with the VESA specification. If it is supplied, an additional four connections are used at the bottom of the connector. The digital signals are transferred on the 24 pins shown.

Table 22.2 shows the connections in the DVI connector. The two italicized rows in the center are used for link 1 and are not used for HDTV applications. Display and clock data are transmitted as balanced signals. Each data channel is individually shielded, with the shield ground being shared between link 0 and link 1. The DDC clock and data are used to identify the type of display, controlled by the VESA *data display channel* specification. Negotiations carried over this link allow the host and the display to agree on operating parameters. The hot plug detect on pin 16 places a ground on the host pin 16 when a display is plugged in. It allows a user to plug in a display after the host is powered up. The ground generates an action by the host to determine the type and capabilities of the monitor that has just been connected.

Figure 22.5(c) illustrates the method used for clock and data signal transmission between the host and the display. This is a balanced transmission technique using terminating resistors at the receiver and switched current sources at the transmitter. Note the similarities between this method of data transmission and balanced audio transmission as shown in Section 8.9.2. Balanced transmission is known for reliable transmission while generating low radiated emissions. Data actuates the switch in the transmitter, which alternately connects the data + and data− lines to the current sink, a circuit that pulls 10 mA to ground at all times.

Table 22.2 DVI connector pin functions

Pin	*Signal Assignment*	*Pin*	*Signal Assignment*	*Pin*	*Signal Assignment*
1	TMDS data 2−	9	TMDS data 1−	17	TMDS data 0−
2	TMDS data 2+	10	TMDS data 1+	18	TMDS data 0+
3	TMDS data 2/4 shield	11	TMDS data 1/3 shield	19	TMDS data 0/5 shield
4	*TMDS data* 4−	12	*TMDS data* 3−	20	*TMDS data* 5−
5	*TMDS data* 4+	13	*TMDS data* 3+	21	*TMDS data* 5+
6	DDC clock	14	+5 V power	22	TDMS clock shield
7	DDC data	15	+5 V ground	23	TDMS clock +
8	N.C.	16	Hot plug detect	24	TDMS clock −

When the switch is open for the data+ line, it is closed for the data− line, so current is pulled through R1, creating a nominal half-volt drop in R1. Since no current is being pulled through R2, the data+ line is at the supply voltage, nominally 3.3 volts. When the switch changes states with data, the opposite occurs. Figure 22.5(d) illustrates the waveform on data+ and data−. The voltage varies between +3.3 V and +2.8 V, with the two data lines at opposite polarity at all times. Thus, the total nominal swing between the two data lines is 1 V peak-to-peak. A differential amplifier in the receiver converts the differential signal to a single-ended signal for use within the receiver.

Since all signals are balanced with respect to ground, and assuming the two wires carrying each signal in the cable are twisted, then very little signal can radiate. There are rather wide tolerances on the voltages and currents shown, so you may measure a significantly different voltage swing in a real system. There is no absolute maximum specified length of the cable; indeed, it would depend on the clocking speed. For test purposes, the cable is to be 5 feet long. Nominal length to connect the cable between a STT and a monitor might be 6–12 feet; there is some thought that at least at lower speeds the cable could work to about 35 feet. In contrast to the IEEE 1394 interconnect, DVI is not intended for bus applications. It is not possible under the existing standard to transfer individual picture components and have the display put them together, so it is necessary for the multimedia processor to assemble the complete picture before sending it to the monitor.

Copy protection for DVI is described in Section 22.6.4.

1394a

This interface is specified for standard-definition digital displays and for recording. It is a serial bus designed, among other things, for interconnecting a cluster of entertainment devices, such as an STT, a display, a camera, a multichannel audio system, and whatever else the user wishes to connect. It is designed to transmit control information as well as program audio and video, and it can transmit various elements of the final picture separately, relying on the display to put them together. Whereas DVI is a point-to-point link intended to connect a video source to a video display, the IEEE 1394 link includes additional control features and is intended to be connected in bus form, interconnecting multiple devices. MPEG compressed video is transmitted on IEEE 1394a links, as opposed to the uncompressed video transmitted on DVI links.

IEEE 1394 is also known commercially as *FireWire*™ or *i.LINK*™. The Consumer Electronics Association (CEA) has specified compatibility standards for using IEEE 1394 in video interfaces, EIA-775-A, *DTV 1394 Interface Specification*.[8] Equipment conforming to this standard bears the logo for *DTV Link*™. Products bearing this logo also conform to the profiles of EIA/CEA-849-A, *Application Profiles for EIA-775-A Compliant DTVs*. The receiver or display has the capability of decoding an MPEG2 datastream, so the information passed over

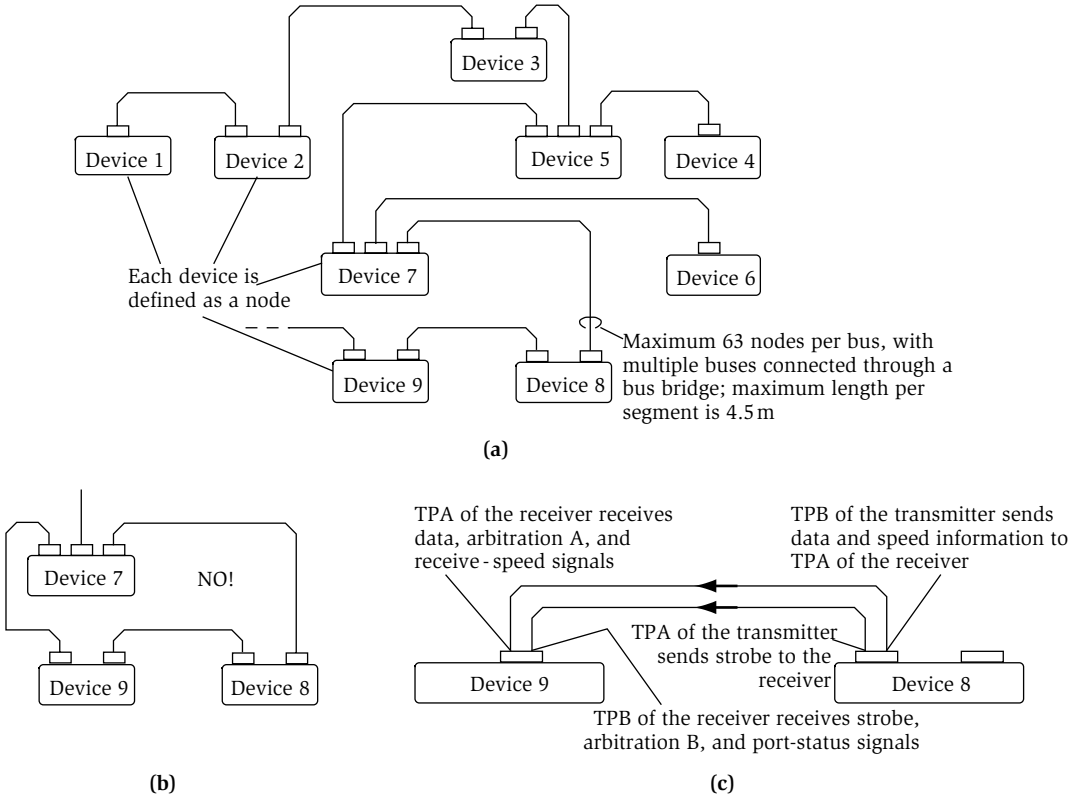

Device 3
Device 1
Device 2
Device 5
Device 4
Each device is defined as a node
Device 7
Device 6
Maximum 63 nodes per bus, with multiple buses connected through a bus bridge; maximum length per segment is 4.5 m
Device 9
Device 8
(a)
Device 7
NO!
Device 9
Device 8
(b)
TPA of the receiver receives data, arbitration A, and receive-speed signals
TPB of the transmitter sends data and speed information to TPA of the receiver
TPA of the transmitter sends strobe to the receiver
Device 9
Device 8
TPB of the receiver receives strobe, arbitration B, and port-status signals
(c)

Figure 22.6 IEEE 1394 connection system. (a) Connecting multiple devices. (b) Incorrect connection. (c) Use of A and B twisted pair.

an IEEE 1394 link is not decompressed in the host. Furthermore, the host can pass a bit-mapped graphic overlay to the display, and the display will overlay it on the picture once it decompresses it. Thus, it is not necessary for the 1394-delivered picture to pass through the multimedia processor of Figure 22.1.[9] On the other hand, the multimedia processor does need to provide the graphics information to the 1394 interface so that it can be passed to the host.

Figure 22.6 illustrates several aspects of the IEEE 1394 bus as used in entertainment applications. A number of devices, called *nodes*, are connected together. The data rate is up to 400 Mb/s today, with higher data rates being developed. (Actual wire data rates are 98.304 Mb/s, 196.608 Mb/s, and 393.216 Mb/s. Conventionally the data rates are rounded to the nearest hundred Mb/s.) Termination of each connection is essential at these rates, so each device must regenerate the signal to pass it on to the next device, even if it is not otherwise involved in a

particular transaction. A maximum of 63 nodes may be connected together, as shown in Figure 22.6(a). If more nodes are desired, a bus bridge may be used to connect them. The maximum length of each cable segment is loosely specified to be 4.5 m (nearly 15 f), though this is not necessarily considered to be a limit. Advanced versions of the IEEE 1394 specification allow longer-distance operation on twisted-pair and fiber-optic cable.

Figure 22.6(b) illustrates an illegal connection of nodes. Here the three devices are interconnected by a loop, so there is more than one path for a data signal to take. This connection will cause the IEEE 1394 interconnection to fail. As long as a loop connection is not made, the system can automatically determine what devices are connected and can determine the capabilities of all other devices in the network. Connections may be daisy-chained and they may be branched, but they may not be looped.

As illustrated in Figure 22.6(c), the IEEE 1394 interface consists of two circuits that are crossed in the cable. Each circuit consists of a twisted-pair controlled-impedance cable, giving rise to the nomenclature of TPA and TPB. In the cable, TBA on one end is connected to TPB on the other, and vice versa. Both pair are used for data communication in both directions in a half-duplex communications protocol. (Half-duplex means that communication is carried in both directions but at different times.) Among other information, *data* is transmitted on TPB and received on TPA, and a *strobe* is transmitted on TPA and received on TPB. The transmission system is nonreturn-to-zero (NRZ) data in which each bit is represented by only one logic level. To aid in clocking, the *strobe* line makes a transition every time two consecutive bits are the same. This ensures one transition for every bit, which aids in clock recovery at the receiver.

Figure 22.7 illustrates the IEEE 1394 cable in more detail. The original 1394 specification, adopted in 1995, called out a six-conductor connector, which accommodated the two twisted pair (TPA and TPB), as well as a power source (8–40 Vdc at 1.5 A maximum). When 1394a was adopted in 2000, a four-pin connector was used, with the power requirement being dropped. This allowed a more compact connector, with each device responsible for its own power. Only the four-pin connector is specified in OpenCable.

Figure 22.7(a) illustrates a six-pin to four-pin cable. The shell is used for a shield connection, also called VG on the four-pin connector. This is the only ground connection between the two ends of the cable. Although balanced transmission is used, it is still necessary to maintain an approximate common ground between the two ends to prevent amplifier saturation and possibly static buildup problems. As shown in the pin function chart, TPA and TPA* (the opposite side of the balanced pair) on pins 5 and 6 of the six-pin connector are crossed to pins 1 and 2, the TPB and TPB* pins on the four-pin connector. Figure 22.7(b) illustrates the more common four-pin to four-pin connector. Note that each twisted pair is shielded, with the shields connected to the shell, VG, at each end of the cable. TPA/A* and TPB/B* are always crossed in the cable.

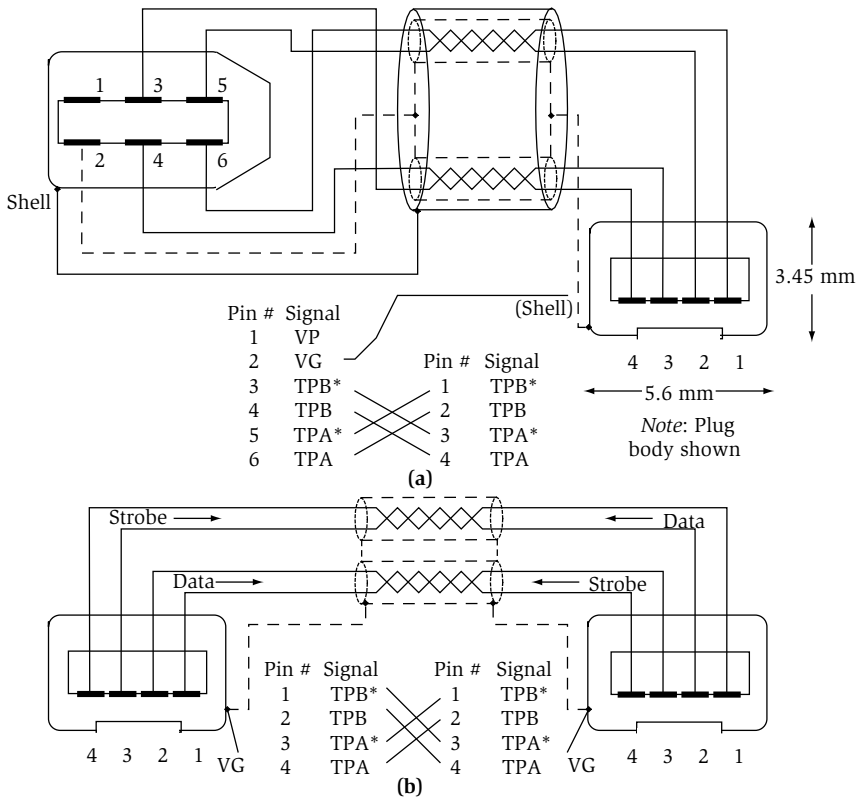


Figure 22.7 IEEE 1394 interconnections. (a) Six-wire to four-wire cable. (b) Four-wire to four-wire cable.

Figure 22.8 illustrates the electrical connections on the two ends of each twisted pair. TPA is illustrated on the left, and the mated TPB pair is on the right. Follow first the transmission from the strobe transmitter (on the left) to the strobe receiver (on the right). The differential impedance of the twisted pair is 110 ohms, or 55 ohms from either side to ground. The source impedance is set by resistors R1 and R2, the series combination of which is placed differentially across TPA and TPA*. The junction of the two is pulled to a bias voltage, $T_P Bias'$. The strobe transmitter is a pair of switched current sources similar to those illustrated in the transmitter of Figure 22.5(c) for DVI (the current source in IEEE 1394 is closer to 2 mA than to 10 mA). By using switched current sources, the source impedance is set by R1 and R2 and is constant regardless of the state of the strobe transmitter.

The signal is transmitted from TPA/TPA* on the left to TPB/TPB* on the right over the IEEE 1394 cable. Resistors R8 and R9 form the differential

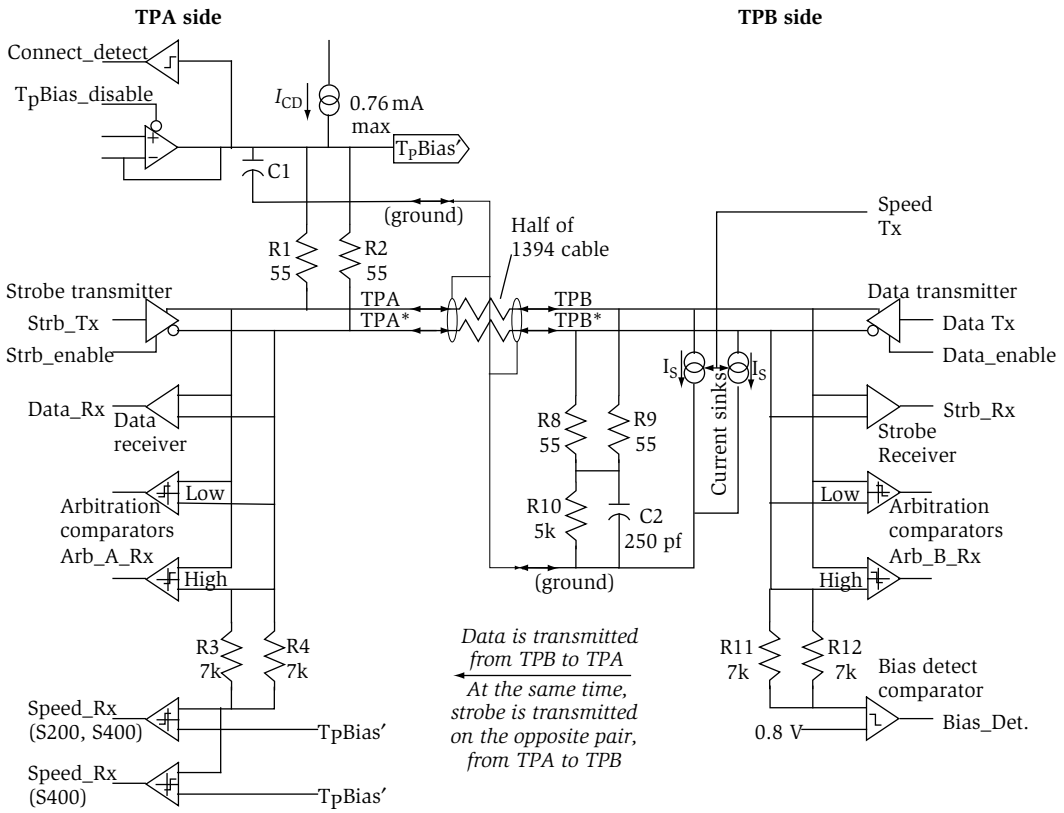


Figure 22.8 IEEE 1394 electrical circuits.

termination at the receive end. The junction of the two is pulled toward ground by R10 and C2. The strobe receiver recovers the logic-level strobe signal. It is a differential receiver, which puts out a logic 1 or a logic 0 depending on which of its two inputs is higher in voltage. The differential voltage amplitude on the signal and strobe lines is between 118 and 265 mV, depending on the speed of transmission.

Since the IEEE 1394 bus can operate at different speeds, it is necessary for two nodes to agree on the speed at which they will transmit and receive. This is done by adjusting the amount of current in the two current sinks I_S on the TPB side. If a node can operate only at 100 Mb/s, then it does not have the current sinks, and the value of I_S is 0. If the node wants to propose (or confirm) operation at 200 Mb/s, then it sets $I_S = 3.5$ mA. For 400 Mb/s operation, $I_S = 10$ mA. The current for these two current sinks comes from TpBias′ on the TPA side through R1 and R2. This current changes the common mode voltage on the twisted pair, with the common mode voltage being detected on the two speed

receivers in the lower left corner of the diagram. The common mode voltage is nominally 1.665 V for 100-Mb/s transmission and 1.438 and 1.03 V, respectively, for 200- and 400-Mb/s transmission.

The arbitration receivers detect the value of differential voltage on the pair during arbitration, the process by which nodes compete for the bus. Arbitration ensures that only one node at a time is transmitting. Upon completion of arbitration, the winning node is able to transmit a packet or initiate a short bus reset. Arbitration signaling is a protocol for the exchange of bidirectional, unclocked signals between nodes during arbitration. The differential voltage may be forced using a combination of the states of the strobe and data transmitters. Two arbitration receivers are needed to distinguish all the possible states of the connection.

The nodes must have some way to know when a new node is added; this is provided by the I_{CD} current source in the upper left. It provides a certain amount of common mode current on the bus when the T_PBias' is disabled. The current creates a voltage drop across R10, which is detected by the bias detect comparator in the lower right. This notifies the existing node that a new node has been connected to it, initiating the discovery process.[10,11]

Two classes of data communication between nodes are supported: asynchronous and isochronous. *Asynchronous* transmissions occur whenever they occur, not synchronized to any event and not at any particular interval. They are always acknowledged. *Isochronous* communications occur at regular intervals. An example of asynchronous traffic might be a control message telling a node to adjust its volume or to change a channel. An example of isochronous traffic would be a packet of video or audio information being delivered from an STT to a monitor. These packets must be delivered promptly and at a regular interval if good performance is to be maintained. IEEE 1394 provides for isochronous cycles 8,000 times a second. There is often little point in acknowledging an isochronous packet, because by the time a missed packet could be retransmitted, it is too late to use it. Isochronous cycles take priority over asynchronous traffic, but the protocol guarantees that a minimum portion of the bus time is reserved for asynchronous data transfers. An ongoing isochronous communication between nodes is referred to as a *channel*. Once a channel has been established, the requesting node is guaranteed to have the requested amount of bus time for that channel every isochronous cycle. Only one node may send data on a particular channel, but several nodes may receive the data. A single node may have multiple channels allocated, and additional channels may be added as long as isochronous capacity is available.[12]

Digital Audio—S/P DIF

Some people actually do try to make a word (spe-dif) out of this acronym. It stands for the *Sony/Philips digital interface*, a serial digital audio interface that can be used to carry either uncompressed stereo or compressed (that is, non-PCM data)

Dolby Digital (AC-3) with up to 5.1 discrete channels of surround sound. (AC-3 audio compression is described in Chapter 3.) The basic specification covering the S/P DIF connection is IEC 60958. It is a 75-ohm coaxial interface using RCA connections. Only one-way communication is supported.

The transmitter supplies $0.5 \pm 20\%$ V p–p into 75 ohms. The receiver must operate with signal levels as low as 0.2 V p–p. The signal is self-clocking at a data rate that is 64 times the *sampling rate* (the rate at which audio data is sampled). As explained in Chapter 6, in order to avoid aliasing, the audio data must be sampled at more than two times the maximum frequency of the signal. If you want to carry audio frequencies up to 20 kHz, you have to sample at greater than 40 ks/s (thousand samples per second). A common sampling rate used in CD audio is 44.1 ks/s, whereas for certain other high-quality applications, such as DVD and digital broadcast, the audio is commonly sampled at 48 ks/s, resulting in a data rate of 3.072 Mb/s on the S/P DIF interface. A sampling rate of 32 ks/s is also used for some applications. The sampling rate may be varied by up to 12.5% for varispeed applications. A self-clocking coding scheme called *biphase mark* is used in which there is a transition in level at the beginning of each bit cell. If the bit transmitted is a 1, there is an additional transition in the middle of the cell. If the bit is a 0, there is no midcell transition.*

When transmitting uncompressed two-channel audio, each frame consists of two subframes, one for left channel data and one for right channel data. Up to 24 audio bits are supported per subframe, with each subframe a total of 32 bits long. (Conventional CD audio data is limited to 16-bit word length and is sampled at 44.1 ks/s.) The additional bits are used for synchronization, auxiliary data, audio sample validity, a user bit, audio channel status, and subframe parity. Two subframes make up one frame, which thus includes one data sample of each of the left and right channels.

A two-channel Dolby Digital (AC-3) decoder will usually be available in the host in order to permit the recovery of digitally tiered audio programming. In some cases, programmers may offer discrete multichannel programming (that is, 5.1), and the host audio decoder must be capable of presenting this program either in stereo via the line level outputs or as a monophonic down-mix via the RF remodulator while the host is tuned to digital programming. This task is accomplished by down-mixing the multichannel audio program using a set of rules carried within the Dolby Digital bitstream using metadata. However, in many if not all cases, the host may default to placing Dolby Digital (AC-3) data on the digital audio interface (the data format is specified in IEC 61937) for use in home theater

---

* Biphase mark encoding is similar to, but not to be confused with, Manchester encoding, another self-clocking mechanism in which a transition always occurs in the center of the bit. The first half of the bit time contains the bit being transmitted; the second half of the bit time contains the complement of the bit being transmitted. Manchester encoding is described in Chapter 4.

systems. The receiving device will then convert the Dolby Digital (AC-3) encoded signal to discrete 5.1 channel audio. IEC 61937 defines the format for carrying nonlinear PCM-encoded audio bitstreams over the IEC 60958 interface. To summarize this format, the Dolby Digital (AC-3) data is formed into a sequence of data bursts, where each data burst is made up of a 64-bit burst preamble (Pa, Pb, Pc, and Pd) followed by the burst payload itself (that is, the encoded audio data). The burst preamble consists of a sync word (Pa and Pb), information about the burst payload (Pc), a bitstream number (Pc), and a burst length code (Pd). The complete data burst is then broken up into 16-bit chunks and placed into the audio payload area of the IEC 60958 subframe, in time slots 12 to 27. Both subframes are used to simultaneously carry 32 bits of data. This allows IEC 60958 to convey nonlinear PCM-encoded bitstreams such as Dolby Digital (AC-3) to home theater systems. For more complete information, refer to IEC 61937.[13]

22.5.3 Basic Concepts of Copy Protection

Before we discuss specific copy protection measures, we need to establish the several basic concepts that different copy protection methods have in common. These include the concept of encryption, public–private key pair, and a certificate. Hang in through this discussion, because after we cover this material, we'll show how these concepts are used in real host applications.

Encryption

Encryption (called *scrambling* in the context of cable TV program denial) is the process of rendering a packet of data unintelligible but capable of being returned to its intelligible, or *unencrypted*, original form by some process. Encryption uses a key or key pair known by the sender and the receiver. The keys may be identical (a *symmetrical* key system) or different (*asymmetrical* keys). A *key* in this context is a binary sequence used to encrypt another binary sequence. That is, I take so-called *plaintext*, which could, for example, be a packet of data representing a portion of a picture I am sending from a host to a display and/or a recording device. I need to encrypt the packet so that it cannot be read by another device, which might usurp the information for a purpose not intended by the owner of the content. In essence, I combine the plaintext with a key: some piece of data that renders the plaintext unable to be read (that is, encrypted). The intended receiver, maybe a video display, knows how to decrypt the packet, returning it to the original plaintext that can be displayed. That is, the intended receiver needs a *decryption key*, a piece of information used to recover the original information.

Figure 22.9 illustrates a basic encryption system. Suppose Bob wants to send a message to Alice. (Let "Bob" be the host and "Alice" be the display. For some unknown reason, the people who study encryption, cryptographers,

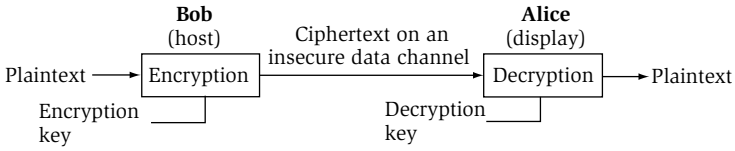


Figure 22.9 Basic data encryption.

traditionally refer to communication between Bob and Alice. We don't know what happened to Carol and Ted.) Bob has a piece of plaintext, maybe a packet of video display data, that he is ready to send to Alice, the display. But he must send it over an insecure channel, the cable between the two. This cable is insecure because it is possible for the subscriber to connect it to some pirate device, maybe to copy the program to a DVD that can be duplicated and given to people who have not paid for the program. Or the subscriber might do something else with the program that is not authorized by the content owner.

In order to allow communication over this insecure channel, the packet is encrypted before being sent to the receiver. A simple form of symmetric encryption is the use of an exclusive OR logic gate. The exclusive OR, XOR, operation is used to add two binary numbers; here it is used to encrypt data. Bit by bit, the data in the packet is presented to the XOR gate. An XOR gate puts out the input bit if the encryption key is a logic 0, and it puts out the opposite bit if the encryption key is 1. Table 22.3 shows the logical operation. When the signal reaches the receiver (Alice), the same encryption bit sequence is used to recover the original information. In order for this to happen, the sender and the receiver have to share a secret, the encryption key. Exchanging, or agreeing on, the key is a major issue in cryptography. (Note that in some of the protocols discussed later, the sender and the receiver don't have the same key, but rather they have related keys.) Somehow Bob and Alice must agree on a secret key or a pair of keys, but the only channel they have for communicating is an insecure one.

Table 22.3 XOR logic

Plaintext	*Key*	*Out*
0	0	0
0	1	1
1	0	1
1	1	0

Public–Private Keys

Exclusive OR encryption as shown in Table 22.3 is representative of simple symmetric key systems. There are also asymmetric key systems, some of which rely on an exponentiation operation rather than an exclusive OR operation. They tend to be much more computationally complex, but they yield some nice advantages, such as simplifying the key exchange problem.[14] Asymmetric keys are used in public–private key pairs.

Figure 22.10 illustrates a basic public–private key system. Suppose Bob needs to send that video packet to Alice. First Alice selects (or in some cases has stored at manufacture) a random number that only she knows. This random number is her private key. From that she can compute a public key (step 2) by applying certain prescribed mathematical operations to that private key. One method of computing a public key from a private key is given later in the description of Diffie–Hellman key exchange. The operations are such that the calculation she does is straightforward, but the inverse calculation is extremely difficult. We show a real example later in the chapter. Alice gives Bob her public key, which Bob uses to encrypt the message. Alice can decrypt it using her private key, which was not sent over the channel. It is not proved here that you can derive a public–private key pair that works this way, but since the mathematics is extremely complex, we must ask you to accept this as fact and refer you to texts that expand on the subject.[15,16]

An eavesdropper who learns Alice's public key still cannot decrypt the message because the keys being used are asymmetrical: Alice's public key encrypted the message, but it cannot decrypt it. Only Alice knows her private key, the only one that can decrypt the message. So the message is relatively safe being passed on the same insecure data channel used to send the encrypted data.

Either the private key can be used for encryption and the public key for decryption or the other way around: the public key can be used for encryption

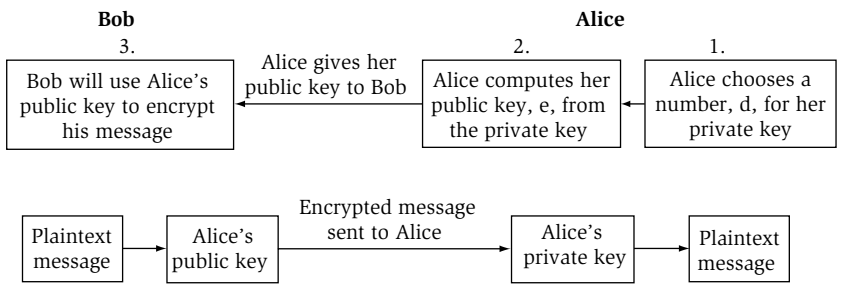

Bob
Alice
3.
2.
1.
Bob will use Alice's public key to encrypt his message
Alice gives her public key to Bob
Alice computes her public key, e, from the private key
Alice chooses a number, d, for her private key
Plaintext message
Alice's public key
Encrypted message sent to Alice
Alice's private key
Plaintext message

Figure 22.10 Public–private key exchange.

and the private for decryption. The same key cannot be used for both encryption and decryption.

Digital Signatures and Certificates

A problem remains. How does Bob know for sure that it is Alice who is sending her public key and not an interloper? For example, how does Bob, the host, know he is not sending video to a pirate box that invented a public–private key pair? After all, the method of calculating the public key from the private key is public knowledge. Maybe "Alice" is a pirate box that will decrypt the picture and record a DVD. This possibility is avoided by the use of digital certificates. There are many, but a common one is the X.509 certificate, promulgated by the International Telecommunications Union (ITU).

The first important concept in certification is not technical: There must be a trusted party to issue certificates. That party, called a *certificate authority* (CA) is a third party that both parties to the data exchange have decided to trust. There are companies set up to be this CA, and there are organizations that act as their own CA for specific purposes. If Bob and Alice wish to communicate, their manufacturers must get them certificates from a CA. Bob cannot send pictures to Alice unless he knows for sure that "she" is a legitimate display who will not usurp the video for a nefarious purpose. Alice cannot get a certificate until the CA has determined, using whatever steps it deems appropriate, that Alice is legitimate and will not violate the usage conditions established by the content provider.

So Alice's manufacturer must go to the appropriate CA and receive a digital certificate before he can build Alice. The digital certificate is then stored in Alice upon manufacture. Each and every copy of Alice, that is, each and every display that is built, will have its own certificate individually prepared by the CA. Alice's manufacturer sends data for the certificate to the CA, which encrypts portions with its private key and returns the finished certificate to Alice's manufacturer. Fortunately, once the credibility of Alice's manufacturer has been manually established, the certificates may be constructed automatically.

Figure 22.11 shows the current version, 3, of the X.509 certificate, the information stored in Alice, provided by the CA. Even Alice's manufacturer cannot construct the certificate, because it involves the CA's private key. The certificate must be individually prepared by the CA based on information supplied by Alice's manufacturer. The certificate is a public document. It does not include enough information for anyone to decrypt data. Rather, it is used only for the purpose of identifying the intended receiver of information and also to convey Alice's public key.

The *version* section of the certificate identifies the format of the certificate (version 3 is current as this is written). The *serial number* is a number unique within the CA by which it can identify the certificate. The *algorithm identifier* conveys the encryption algorithm used by the CA (often the RSA algorithm,

Version	Certificate version
Serial number	Unique number within the CA
Algorithm identifier: - Algorithm - Parameters	Identifies algorithm used to sign the certificate and any needed parameters
Issuer	Name of the CA
Period of validity: - Not before (date) - Not after (date)	Certificate is good only between these dates
Subject	Name of the user
Subject's public key: - Algorithm - Parameters - Public key	Algorithm name and necessary parameters, plus the public key
Issuer unique ID	
Subject unique ID	
Extension	
Signature	CA's signature

Figure 22.11 X.509 certificate.

named for its inventors, Rivest, Shamir, and Adleman, is used). There are some public parameters used in specifying the public key from the private key, and these are included with the algorithm identifier. (In early versions of the standard, this field is called the *signature*, but this is a misnomer.)

The next section of the certificate is the name of the CA. The *period of validity* specifies the beginning and ending dates that the certificate is good. Outside of these dates, the certificate is invalid and may not be used. *Subject* is the name of the user. The next section bears the subject's public key. It names the algorithm being used (often RSA) and any parameters associated with the public key. Finally, this section provides the public key that will be used by the sender to encrypt data.

The *issuer unique ID*, the *subject unique ID*, and the *extension* were recently added and may or may not be used. These are fields that may be used by the issuer or user as desired. The last section, the *signature*, is very important. The signature is the subject's (Alice in our example) public key plus other certificate data, encrypted with the CA's private key. This is the reason that the CA must provide individual certificates for each and every device made by a manufacturer. No one, and we mean no one, except the CA must know the CA's private key. If this private key is ever compromised, then all certificates issued by the CA are rendered invalid.[17]

During an initial discovery process, Alice sends her certificate to Bob. Bob must know the CA's public key. He can know it by virtue of its having been programmed into him during his manufacture or by obtaining it later. Bob decrypts Alice's public key and other information from the signature, using the CA's public key. He compares the public key and other information with the same data sent as plaintext in the signature. If the two sets of information match, then he must be talking to a legitimate Alice, because a pirate would not be able to get a certificate from the CA. Bob now has Alice's public key, so he can send data to her, encrypted with her public key.

The certificate is public knowledge. Can an interloper steal Alice's certificate? Yes. It wouldn't do any good, because it does not contain her private key, but the interloper could steal it. As an additional verification that there is a legitimate "Alice" out there, Bob can issue a challenge. He sends Alice a random number called a *nonce*, which she encrypts with her private key and sends back to Bob. Bob decrypts it with Alice's public key, which he received as part of her certificate. If the decrypted number matches the number Bob originally sent, he knows that the real Alice is out there and not someone who happens to have her certificate.

Can Bob be absolutely sure he is talking only to Alice and not also to an eavesdropper? No, but it doesn't matter. There could be a so-called *man in the middle*. For example, a man-in-the-middle attack could happen if the subscriber has connected a legitimate host (Bob) to an illegitimate device designed to capture the video on its way to Alice. The illegitimate device might then be connected to the real Alice, the legitimate display. The man in the middle passes data between Bob and Alice, but he eavesdrops, or hears all that is going on. When Alice sends her certificate to Bob, the man in the middle passes it through to Bob, who recognizes it as Alice's legitimate certificate. But it will do the man in the middle no good to have the certificate. He still doesn't have Alice's private key, which he would need in order to decrypt what Bob sends to Alice. Because Bob got Alice's public key in the certificate, he knows that he will be encrypting data for a legitimate display, regardless of who else is out there.

So Bob knows that the real Alice is out there because no one except the real Alice would have her certificate *and* be able to pass the challenge he sent. He knows the certificate is real because he was able to match information contained therein using the CA's public key to decrypt part of the certificate and to match it to the same information sent unencrypted. Only the real CA could have signed the certificate. And both sides have placed their trust in the CA. Bob can safely send data to Alice, knowing that the real Alice is receiving it. If anyone else is out there, the data will be useless to him.

In some instances the CA is designated by the originator of the standard utilizing the certificate. For example, for DVI, HDCP is the encryption used. It specifies that Digital Content Protection, LLC, an activity set up by the Digital Display Working Group that defined the specification, is the CA.[18] Anyone desiring to make a device that either sends video to or receives video from a

device conforming to the DVI specification must apply to them for licensing. For the IEEE 1394 interface, licensing of the DTCP copy protection algorithm and certification is provided by the Digital Transmission Licensing Administrator, DTLA.[19]

In the present application, a certain CA is usually specified. However, in the more general case it is possible to have two devices that were certified by different CAs. Certification may still take place through the use of a chain of certificates.

Certificate Revocation

From time to time it is possible that some device possessing a valid certificate may need to have its certificate revoked. This might happen if the CA's private key becomes known or, perhaps more likely, if it is discovered that the device has been modified into a pirate device. Many systems have a mechanism for disseminating to devices that they should not honor a certain certificate. One method of disseminating the information is that as soon as the breach of security is recognized, future content will have embedded in it the identity of the certificate that must be revoked. As that content is disseminated to different devices (hosts, legitimate digital VCRs, etc.), each remembers the revoked certificate and passes that information to devices to which it is, or becomes, attached. In this way, sooner or later most devices that may connect to the device with the revoked certificate will learn of the revocation.[20] In some cases, discussed later, the host must check with a database to confirm the validity of a certificate.

Key Length, Security, and Computational Complexity

One attack on an encryption key is called an *exhaustive search*. Encryption keys are simply long strings of 1s and 0s. In an exhaustive search, a would-be pirate simply tries all possible key combinations (sets of 1s and 0s) until the output makes sense. Often this can be automated, because incorrect keys yield results that can be distinguished, by computer, from logical results. The longer the key, the more difficult is an exhaustive search. If a key has a length (number of binary characters) of n, then there are 2^n combinations that must be tried. A short key might have only 24 bits, so there are a maximum of $2^{24} = 16.78$ million combinations that might work. If a computer can try 1 million keys per second, then the maximum time required to crack the encryption is 16.78 million divided by 1 million, or about 17 seconds. On the other hand, a 56-bit key could take 2300 years to crack at the same rate. Keys are used that have up to 256 bits, which could take 3.7×10^{63} years to break at the stated rate of trying keys. Average times, though, would only be half of these numbers.

Sometimes *churning keys* are used to improve security. This technique simply means that the key is changed fairly often. Each time a key is changed,

the pirate would have to begin again to break it. So a pirate may capture a scrambled program, but to unscramble it he would have to break a series of keys that might be changed every few seconds to every few minutes.

Symmetric Algorithms

Many asymmetric algorithms, such as the RSA algorithm used in a number of systems, involve taking a block of plaintext (treated as a binary number) to the power of the key, using some modulus. The computational power required as the key length increases becomes prohibitive for encrypting and decrypting video streams. Computational power may be traded off for time; but when you are transferring digital video, whether compressed or not, you don't have a lot of time to process the signal. Thus, many times a good asymmetric algorithm, such as RSA, will be used for certificate verification and/or key exchange, which are only done periodically. Then a symmetric algorithm will be used for actual data encryption. In the DTLA nomenclature, this symmetric key is called the *content channel key*. The key may be long, but the way it is used reduces the computational complexity to something manageable at the requisite speed.

Often the RSA algorithm is used for certificate validation. Then a symmetric key is agreed upon for the actual information scrambling and descrambling. The symmetric key is used in some manner to generate the actual encryption string, which in some encryption techniques is exclusive ORed with the plaintext to create the encrypted data, or ciphertext.

Figure 22.12 illustrates a simple four-bit shift register used to generate ciphertext.[21] Real shift registers are a lot longer, and multiple shift registers may be used. The shift register is clocked at the same rate as the plaintext data is clocked. Each time the shift register is clocked, the bits in each position shift one place to the right. (The clock is a signal that synchronizes operations dealing

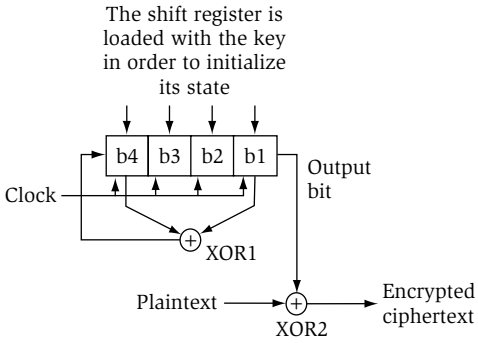


Figure 22.12 Linear feedback shift register used to generate ciphertext.

with delivery of the video.) The shift register is a series of flip-flops. On a clock cycle, the bit (1 or 0) in b4 is shifted to b3, the bit in b3 is shifted to b2, and so on. The bit in b1 is supplied to XOR2, which exclusive ORs it with the plaintext bit presented on the same clock cycle.

This type of shift register is called a *linear feedback shift register* (LFSR). The linear feedback is supplied by XOR1. It XORs the output of two bits, in this case, b4 and b1, feeding the result back as the input to the shift register, which is loaded into b4 on the following clock cycle. The bit sequence (a pseudo-random stream) obtained from a properly designed n-bit LFSR is $2^n - 1$ states (clock cycles) long before it repeats. It can be initialized with the secret symmetric key, which loads the starting state into the LFSR. A listener will not be able to duplicate the LFSR initialization without knowing the key used to initialize it.

Many other methods of generating the encrypting code are known. A different LFSR may be used for each bit of a parallel encryption system. The LFSRs may be initialized with the same key or different portions of the same key. Nonlinear generators may be used, in which AND functions combine the outputs of various stages to form the output. The outputs of several LFSRs may be combined to form the encryption signal.

Diffie–Hellman Key Exchange Protocol

Though by no means the only useful key exchange protocol, the Diffie–Hellman (D–H) protocol is used in quite a few systems. D–H permits both parties to an exchange (Bob and Alice) to supply part of the secret key. It is done in such a way that the entire key is not sent across the unsecured channel. Thus, a snooper is not going to gain the information needed to steal the secret key. The protocol proceeds as follows.

1. Alice chooses a random large integer x and sends Bob $X = g^x \bmod n$, where g and n are D–H parameters that must be passed between the two parties. For example, suppose D–H is used to agree on a key in an X.509 certificate. In the "subject's public key" portion of Figure 22.11, g and n would be passed as parameters, and D–H would be identified as the algorithm. X is the public key. The parameter n must be large and must meet certain criteria, but g may be a single-digit number.

   These parameters are not secret, and an eavesdropper will not know anything useful if he learns them. In some closed systems, g and n may be preprogrammed during manufacturing. The parameter that must be kept secret is x, and, as we shall see, Alice never divulges it in any form. The modulus, n, represents the maximum value the result can take (plus 1). It is the remainder after a division by the modulus. To calculate $X = g^x \bmod n$, take the remainder from the operation g^x/n.

2. Bob chooses a random large integer y and sends Alice $Y = g^y \bmod n$.
3. Alice computes $k = Y^x \bmod n$. Recall that Y was sent to her by Bob and that x is a secret parameter that only Alice knows.
4. Bob computes $k' = X^y \bmod n$. He got X from Alice and knows y. Now, from the foregoing (and not writing the modulus), $k' = X^y = g^{xy}$. Also, $k = Y^x = g^{xy}$, so k and k' are equal. Thus, Alice and Bob have the same secret key, and at no time did all the information required to know the secret key pass over the insecure link.

$k(= k')$ is the symmetrical key applied to the LFSR in foregoing. It is possible to derive a D–H key among three or more parties by expanding on the preceding process.[22]

In order for an eavesdropper to learn the secret key, he would have to derive x and y from X and Y, the two public keys that are passed over the link. Computing x and y is not feasible. And without them, knowing k is not possible. One might attempt to do an exhaustive search, but it will take many lifetimes to do an exhaustive search if x and y are large enough.

DES Encryption Algorithm

There is yet one more encryption algorithm we need to discuss, because it is used in handling digital video. This is the data encryption standard, or DES (usually pronounced "dez"). DES is based on work done at IBM in the early 1970s. The National Institute of Standards and Technology (NIST, formerly NBS, the National Bureau of Standards) established it as a national standard in 1976. The American National Standards Institute approved DES as a private-sector standard in 1981 (ANSI X3.92). It has been reaffirmed as a U.S. standard several times since then, though it is now no longer approved for federal information applications, having been replaced by AES in 2001. Just because DES is no longer the approved algorithm does not mean that it is broken, just that it has been replaced for federal information applications. Special-purpose DES chips are available.

DES is a 56-bit-key symmetrical algorithm, meaning the same key is used for encryption and decryption. Though complex, it uses only simple operations, such as exclusive OR, permutation (changing the order of bits), and shifting of bits. There are four modes of operation: electronic code book (ECB), cipher block chaining (CBC), output feedback (OFB), and cipher feedback (CFB). We shall describe the complete algorithm, though only a portion of it is used in some scrambling systems.

Figure 22.13 illustrates the operations performed on the data. DES is a block cipher; it operates on a block of 64 bits of data, sends that data on, and starts working on a new block of 64 bits. If a block has fewer than 64 bits, additional bits are added (*bit stuffing*) to bring the total to 64 bits. An object of DES is to

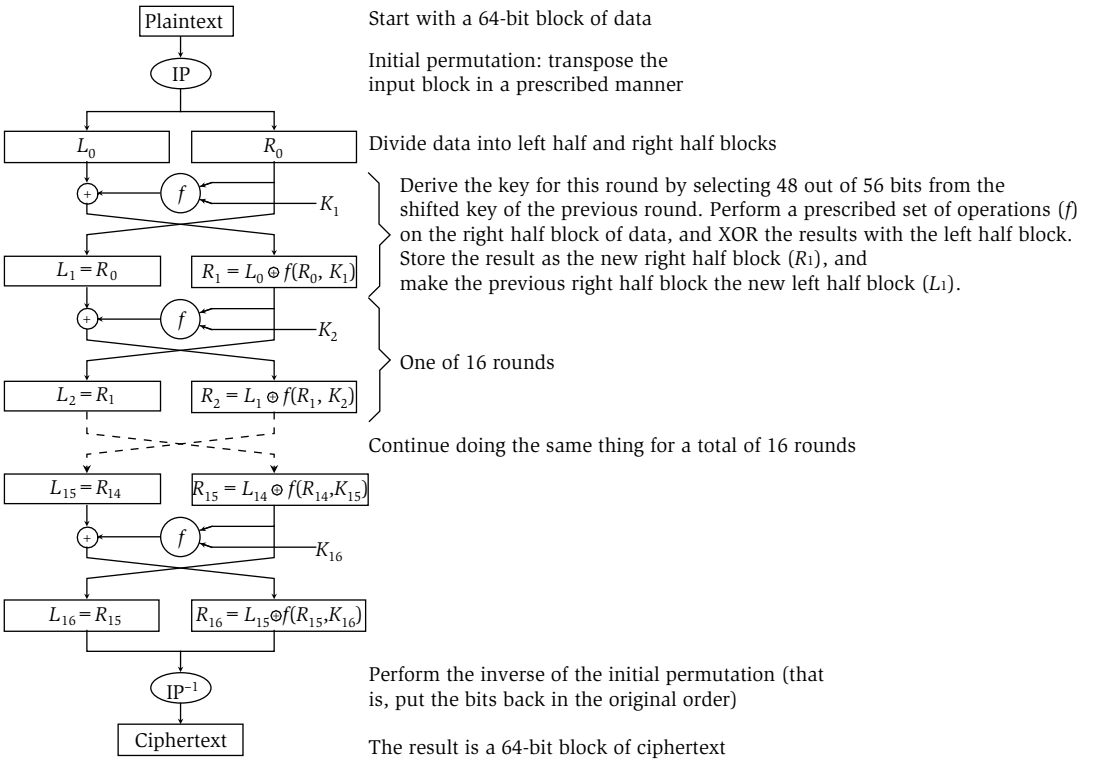

Plaintext
Start with a 64-bit block of data
IP
Initial permutation: transpose the input block in a prescribed manner
L_0
R_0
Divide data into left half and right half blocks
f
K_1
Derive the key for this round by selecting 48 out of 56 bits from the shifted key of the previous round. Perform a prescribed set of operations (f) on the right half block of data, and XOR the results with the left half block. Store the result as the new right half block (R_1), and make the previous right half block the new left half block (L_1).
$L_1 = R_0$
$R_1 = L_0 \oplus f(R_0, K_1)$
f
K_2
One of 16 rounds
$L_2 = R_1$
$R_2 = L_1 \oplus f(R_1, K_2)$
Continue doing the same thing for a total of 16 rounds
$L_{15} = R_{14}$
$R_{15} = L_{14} \oplus f(R_{14}, K_{15})$
f
K_{16}
$L_{16} = R_{15}$
$R_{16} = L_{15} \oplus f(R_{15}, K_{16})$
IP^{-1}
Perform the inverse of the initial permutation (that is, put the bits back in the original order)
Ciphertext
The result is a 64-bit block of ciphertext

Figure 22.13 Operation of DES.

make sure that every bit in the block of ciphertext depends on every bit of the input and every bit of the key. The algorithm is divided into 16 nearly identical *rounds*, each involving all the bits of the data. Only one key is used, but, as we shall see, the bits in the key are changed for each round.

As shown at the top of Figure 22.13, DES starts by performing an *initial permutation* on the block of plaintext. This simply means that the order of the bits is changed in a prescribed manner. The 58th bit becomes the first bit, the 50th bit becomes the second bit, the 42nd bit becomes the third bit, and so on (the referenced text for this section shows all of the permutations involved). The initial permutation has nothing to do with the security of the algorithm. As best anyone can tell today, its primary purpose was to make it easier to load plaintext into a DES chip in byte-sized pieces. Next, the data is divided into a 32-bit left half and a 32-bit right half. Each half will be operated on separately in each round; at the conclusion of each round, the left and right halves are swapped. The left half is referred to as L_n and the right half as R_n, where n is the round number.

For each round, the right half is operated on by a function *f*, which we shall define shorthy. Function *f* involves expanding the data from 32 to 48 bits by duplicating certain bits, XORing the expanded data with 48 bits derived from the key, and performing certain other operations. The result is then XORed with the left half of the bits and the result is stored as the right half for the next round. The right half is stored as the left half for the next round.

After 16 rounds, an inverse of the initial permutation is performed; the resulting 64-bit block is the ciphertext that is transmitted. There is defined a triple DES, in which this algorithm is performed three times using two or three different keys.

Figure 22.14 shows what goes on during each round of DES. On the right we show what happens to the key during each round, and on the left we show what happens to the data during that round. The input key is the key from the previous round or, for the first round, is the agreed-upon key. The actual key is 56 bits long, but it is often expressed as a 64-bit number. The difference is the inclusion of

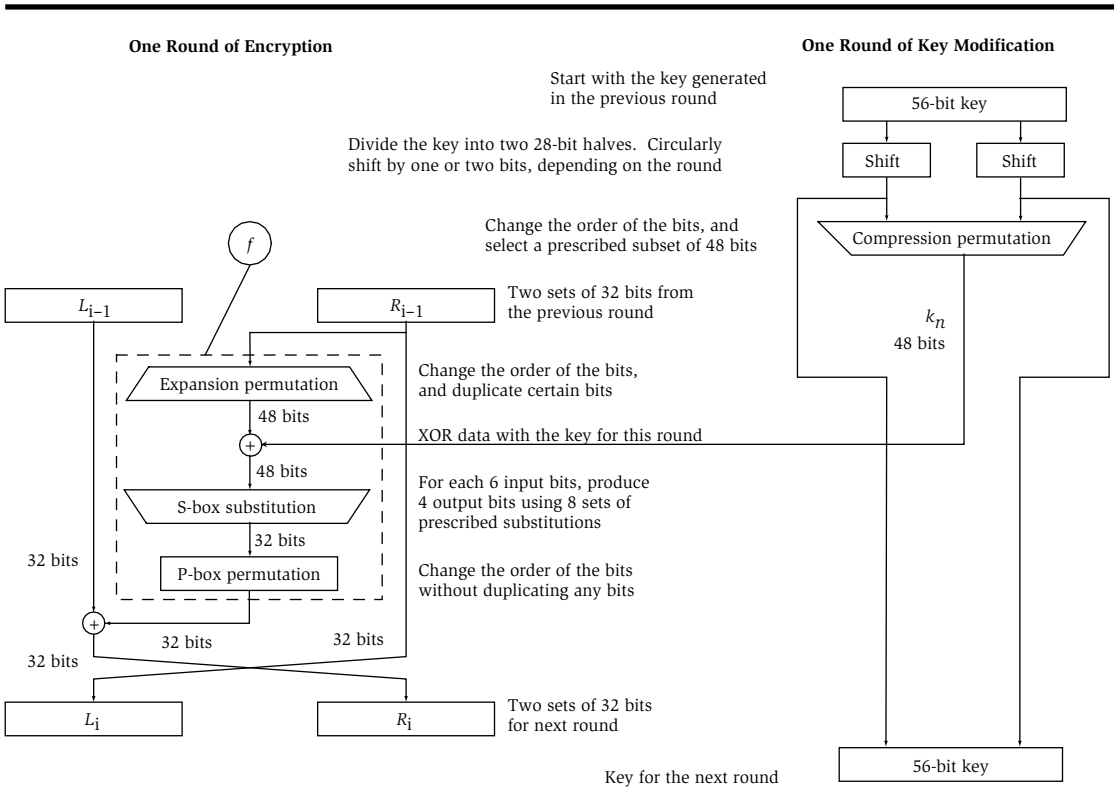

One Round of Encryption
One Round of Key Modification
Start with the key generated in the previous round
56-bit key
Divide the key into two 28-bit halves. Circularly shift by one or two bits, depending on the round
Shift
Shift
Change the order of the bits, and select a prescribed subset of 48 bits
Compression permutation
f
L_{i-1}
R_{i-1}
Two sets of 32 bits from the previous round
k_n
48 bits
Change the order of the bits, and duplicate certain bits
Expansion permutation
48 bits
XOR data with the key for this round
48 bits
For each 6 input bits, produce 4 output bits using 8 sets of prescribed substitutions
S-box substitution
32 bits
32 bits
Change the order of the bits without duplicating any bits
P-box permutation
32 bits
32 bits
32 bits
L_i
R_i
Two sets of 32 bits for next round
Key for the next round
56-bit key

Figure 22.14 Operations performed in one round of DES.

8 parity bits, each derived from 7 bits of the key. The key is divided into halves, and each half is shifted by one or two bits depending on the round. The shifting is done circularly, meaning that bits are shifted in a shift register, with the bit shifted out of one end being shifted into the other end of the shift register.

Next a compression permutation is performed. *Permutation* means that the order of bits is changed according to a prescribed ordering. *Compression* means that certain bits are removed from the key to bring the number of bits remaining down to 48. Again, the bits removed are carefully prescribed. The 48 bits taken from this operation make up the key used in the round. The shifted halves of the key before the compression permutation become the 56-bit key for the next round.

The data encryption performed in this round is shown on the left. The right half block of data (R_{i-1}) becomes the left half for the following round (L_i). Also, the right half is subjected to a set of operations that define the function *f*. What goes on in function *f* is shown inside the dashed lines. The first operation in *f* is an expansion permutation, in which certain bits are repeated to bring the 32 bits up to 48 bits (the same length as the key for the round), and the order of bits is changed. The output of the expansion permutation is XORed with the key.

The next operation is the so-called S-box substitution. A *substitution* is an early form of a cipher. A simple substitution might be to build a table of substitute letters, as shown in Table 22.4. In this simple example, A becomes Z, B becomes P, and so on through the alphabet. A person intercepting such an encrypted message would not be able to make sense out of it unless he had the decoding table. This is the principle behind the S-box substitution in DES. Recall that the input to the S-box substitution is 48 bits, but only 32 of those bits carry the data. The other bits are discarded simultaneously with the substitution. For each six input bits, four output bits are produced according to a prescribed set of eight substitution tables following the same principle as Table 22.4 (the inputs and outputs are sets of bits rather than letters).

The output of the S-box substitution is the input to the P-box permutation. This is a simple permutation of the order of the bits, according to yet another prescribed ordering. The output of the P-box permutation is the output of the function *f* shown in Figure 22.13. It is XORed with the left half bits and placed in the right half block for the next round of the permutation.

Table 22.4 Substitution cipher

In	*Out*
A	Z
B	P
C	J
D	A

The set of actions depicted in Figure 22.14 is repeated through the 16th round, after which the inverse of the initial permutation is performed, as shown in Figure 22.13. The output of the inverse initial permutation is the 64 bits of ciphertext, which are ready to be transmitted, and a new block of 64 bits enters the encryption as the process starts over.

In order to decrypt the data, the same steps are repeated, with the only difference being that the same key is used but the order of processing the key is reversed. Note that we have not proved DES works, nor have we proved the security of DES. For those wishing to study the subject and having several years of free time on their hands, many books have been written on the subject.[23]

22.5.4 Digital Video Copy Protection

Having established the basics of encryption, we can now describe the several copy protection methods used in host devices. Both DVI and the IEEE 1394 interface have associated copy protection mechanisms, and a third protects content coming out of a POD module. (Yet another encryption is applied to the scrambled data transmitted over the cable plant, to be decrypted on the POD module.)

DVI and HDCP

This interface is intended to operate between a host and a display only. Recording the signal is not an appropriate use of the interface. The interface, as described earlier, transfers decoded red, green, and blue pixel information, each serially on three balanced lines. HDCP, the copy protection mechanism, is governed by the specification *high-bandwidth digital copy protection system*, currently revision 1.0, 17 February 2000.[24] Implementation of the specification requires a license from Digital Content Protection, LLC. The system allows for use of *repeaters*, that is, devices that take in the signal and also put it out. A repeater might be a digital video distribution amplifier, or it might be a monitor that supplies an output to another monitor. It is possible to have seven levels of repeaters and as many as 128 devices in total. When repeaters are used, the source (host in our terminology) is made aware of all downstream devices before transmissions commence.

There are three elements of the content protection system. The first is *authentication*, in which the host verifies that all devices connected are legitimate, licensed devices. Shared secrets are established during authentication and form the encryption key for *program transmission*, the second element. Finally, *renewability* allows the host to identify a compromised device and to desist from sending data to it.

Authentication, the first element of the content protection system, also precedes in three steps. It is done slightly differently than what we showed earlier. Each licensed device is given an array of 40 secret keys, each 56 bits

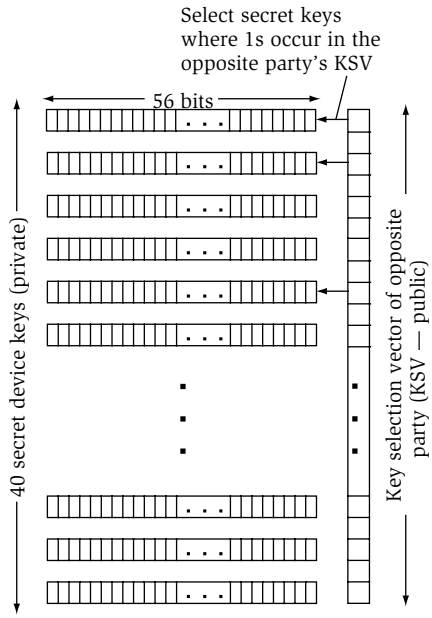


Figure 22.15 HDCP keys.

long. These come from Digital Content Protection, LLC. The device also receives a *key selection vector*, KSV. All are programmed when the device is manufactured. Figure 22.15 shows the secret keys and the KSV of the *opposite* party. The KSV is functionally equivalent to a public key, and the 40 secret keys are just that: secret keys. Authentication is initiated by the host sending an initiation message to the display. This message includes the host's KSV and a 64-bit pseudo-random number, An. The receiver responds with a message including its KSV and a repeater bit if the device is a repeater. The host checks the KSV to make sure it has not been revoked and that it includes 20 1s and 20 0s.

A shared secret is derived as follows, in the first step of authentication. From its secret keys, the host selects the ones corresponding to 1s in the receiver's KSV. All selected secret keys are added together, modulo (mod) 56, producing a 56-bit key. Call this result km in the host and km' in the receiver. The secret keys and the KSVs are chosen such that the host and the receiver will compute $km = km'$, so each end will have the same secret key. If either device has an invalid set of secret keys or a bad KSV, then the two keys will not be the same.

At this point, both devices have several pieces of information in common. They know km, they know the random number generated at the Host, An, and they know the repeater bit. These values are used to generate three values, Ks, M_0,

and R_0. These are generated by a block cipher algorithm defined in the specification. *Ks* is a 56-bit session key. M_0 is a 64-bit secret value used in the second part of the authorization protocol. R_0 is a 16-bit response value that the video receiver returns to the host as in indication as to the success of the authentication exchange. It must be returned within 100 ms. If R_0 returned from the receiver matches that generated by the host, then authentication was successful.

In the second step in the authentication process, the host (transmitter) and any repeaters gather information regarding all the devices that are attached. The host confirms that none of the devices attached are on the revocation list. This step is omitted if there are no repeaters (for instance, there is only one host sending pictures to one display). The step can take up to about five seconds to complete. Finally, the third step occurs in the vertical blanking interval (VBI) about every two seconds while pictures are being transferred. New cipher initialization values are calculated to establish a new 56-bit key for encrypting video content. The integrity of the link is also confirmed.

The second element of the HDCP protocol is encryption of the red, green, and blue data using a pseudo-random sequence generated from secrets exchanged in the authentication process. The pseudo-random sequence is reset frequently. The pseudo-random sequence is exclusive ORed with the data, as shown earlier.

The third and final element of the HDCP protocol is renewability, the process of revocation of devices that have been compromised. The devices are identified by their 40-bit KSV, as described in Figure 22.15. *System renewability messages* (SRMs) containing one or more KSV revocation lists, called *vector revocation lists* (VRLs), are distributed with programs as described above. They are checked by the transmitter (host) when they are available. The SRM carries a digital signature, as described earlier, so that the host can determine that the data did indeed come from Digital Content Protection, LLC.

IEEE 1394 and DTCP

The *digital transmission content protection* (DTCP) standard for secure transmission over IEEE 1394 physical links was developed by five companies — Intel, Matsushita (Panasonic), Toshiba, Sony, and Hitachi — and is thus sometimes called the "5C" content protection scheme. 1394 is intended as a broader-use standard than is DVI, being intended not only for transmission to a display or displays, but also for transfer to recording devices, computers, and whatever other devices you wish. DVI transfers uncompressed (hence very wideband) red, green, and blue data for immediate display; IEEE 1394 transmits compressed (therefore lower bandwidth) data.

DTCP, the associated verification and encryption standard, includes the same elements as does HDCP, with one more. It includes *authentication and key exchange* (AKE), transmission, referred to as *content encryption*, and renewability. It also includes a new element, *copy control information*.

DTCP includes instructions to the receiving device as to how it should treat the data. This is called *copy control information*, CCI, which is embedded in the content. The actual data in the CCI is known as the *encryption mode indicator* (EMI). It may take on any of four states. The material may be used for unlimited copying, copy free, and the opposite state, copy never. Another state is copy once, such as to allow for time shifting a TV program. When the material is played, the EMI is changed to copy no more. These last two values of the EMI are intended for devices with limited computational ability. The associated limited, or restricted, authentication required involves fewer computations than does full authentication.

Obviously, the same concerns apply here as to other systems. Every device licensed to receive material protected by DTCP must obey these rules. Thus, the same need exists for a trusted authority to certify each device. DTCP is administered by the Digital Transmission Licensing Administrator (DTLA), established by the five companies. The DTLA acts as the certificate authority (CA) for conformant devices. Besides the CCI, DTCP includes authentication and key exchange, content encryption, and renewability.

Two levels of authentication and key exchange are specified: Full authentication using public key cryptography as described earlier is used for all types of content. For copy once and copy no more content, a more limited, restricted authentication using common key cryptography is permitted.

For full authentication, the method is similar to, but uses different formats than, that described in Section 22.5.3.[25] A certificate is embedded in each device and is verified as described earlier. Both the source device (supplying the content) and the sink device (receiving the content) check the other device's certificate. A form of Diffie–Hellman key exchange is used for derivation of a secret key used to encrypt program material. Restricted authentication may use a short form of authentication, where the sink device proves that it shares a secret with the source device.[26]

After authentication, the source device (the host) sends an exchange key, encrypted with the authentication key, to the sink device. This exchange key is sent as a seed, a random number generated by the source, from which both devices can compute the decryption key. The key is changed periodically, every 30–120 seconds.

In order to provide for system renewability, *system renewability messages* (SRMs) are sent to each IEEE 1394-compliant device. These may be sent as part of a program stream from a headend, or they may be incorporated into other distribution media, such as DVDs. The SRMs carry identification of any device that has had its certificate revoked. As material containing the identification propagates throughout the nation, devices eventually learn of any certificate revocations.

POD Module Output

The POD module shown in Figure 22.1 is responsible for descrambling the output of the transmission from the headend. POD, an acronym for *point of*

deployment, is the term used for renewable security modules. The POD module is built into a standard PCMCIA (*Personal Computer Memory Card International Association*) card, the familiar card used for modems, Ethernet connections, and many other added functions in portable computers. It uses a 68-pin connector. The POD module output is also copy protected. We can easily explain the first stage of the copy protection algorithm used at this interface. It is the X.509 certificate and encryption algorithm described in Section 22.5.3. The certificate exchange is, however, just the first phase of having the POD module and the host authenticate each other.

In phase 2 of the authentication process, the host and the POD module identify themselves to the headend. The headend holds the DTLA (see preceding subsection) certification revocation list (CRL), showing the identification of all hosts and POD modules that have had their certification revoked. Note the contrast with the general case we showed earlier for revocation of certificates. Previously we showed the CRL being contained with the content. But POD modules and hosts identify themselves back to the headend, which then checks its CRL and will refuse to authorize descrambling if either device is on the CRL.

Identification to the headend is done differently depending on whether a two-way communications path is available and or unavailable. If a communications path is unavailable, then the host displays a message on the monitor, telling the subscriber to call a number and to give the identification numbers of the POD module and the host. The identification numbers are numerical, so it is possible to automate the process of relaying the information to the headend. Also, when the subscriber purchases his or her equipment, it is possible for the seller to perform identification. If two-way communications are available, the POD module can relay the information back to the headend conditional authorization system without subscriber action.

The third phase of authentication consists of computation of a long-term *authentication key* by both the POD module and the host. This authentication key is stored in nonvolatile memory (that is, the content remains even when the device is not powered). It is used to confirm authentication after power has been removed and is reapplied. Upon a subsequent power-up, the POD module and the host confirm that they have the same authentication key; if so, no further action is required before they can begin exchanging programming information.

The complete process of authentication is referred to as a *binding* between the POD module and the host. A POD module may bind to only one host at a time, but the POD module may be moved from one host to another, with the authentication process starting all over when that happens; that is, the POD module will restart the binding procedure. In two-way systems, the process is unseen by the subscriber. But in a one-way system, the subscriber will have to telephone the POD module and host identification numbers to the service provider. Regardless of successful authentication, programs sent in the clear (no scrambling) may be processed.[27]

22.6 Using Two Conditional Access Systems

Operators consider it desirable to be able to use more than one STT system in order to gain possible price and feature advantages. However, each manufacturer would like to use his own conditional access system, to which he ascribes unique benefits. A simplistic approach is to duplicate programming using two scrambling systems. However, this approach doubles bandwidth requirements, an obvious impracticality. In 1997 the so-called *harmony agreement* was negotiated in the industry, whereby one encrypted datastream could be used along with more than one conditional access system.

Besides addressing scrambling, the harmony agreement includes an agreement on the use of open standards for transmitting digital video. These include MPEG-2 video, Dolby Digital audio, MPEG-2 transport, ATSC system information, and ITU-J83B modulation (64- and 256-QAM).[28]

Figure 22.16 illustrates the principle of this harmony agreement. The agreement starts in the headend with clear video (called *plaintext*). This plaintext is encrypted in the common service encryptor, which scrambles the signal using an agreed-upon algorithm, the DES encryption standard described earlier in this chapter. A common (shared) control word, or encryption key, is generated in the control word generator and supplied to the common service encryptor and also to the conditional access system for each vendor whose STTs are being used. The *service encryptor* is the device that scrambles the program data. It is called "common" because the same scrambling is applied for use in both systems.

The control word is also supplied to each vendor's conditional access system, which then distributes that control word as it sees fit. The control word may be encrypted for distribution according to the individual manufacturer's proprietary facility. Control words are embedded in so-called entitlement control messages (ECMs), which are transmitted as part of the MPEG transport stream.[29]

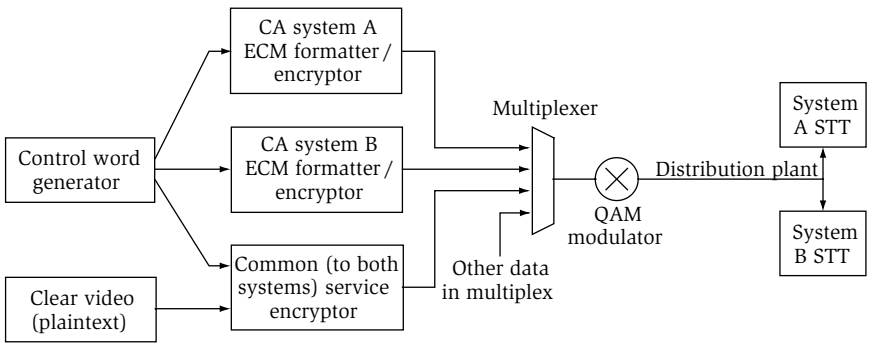


Figure 22.16 Principle of the harmony agreement.

As this book goes to press, another means of sharing conditional access has been demonstrated but not yet commercialized. This method extracts critical data from the MPEG-2 datastream. This data is required for reconstructed video. The critical data is encrypted using any number of encryption methods desired by different vendors. The remaining portion of the MPEG signal is not encrypted. The decoder extracts the critical data encrypted with its encryption system and combines it with the majority of the video data that was not encrypted.

Other means of encrypting the signal are available and are practiced in Europe and other localities.

22.7 Optional Personal Video Recorder

An optional personal video recorder (PVR) function is shown in Figure 22.1. PVR capability allows a subscriber to record and play back material from a hard drive. He or she can use it for time shifting programs and also to generate an "instant replay" of a scene. If the manufacturer wants to allow the PVR to be used with analog signals, the manufacturer must include the optional real-time MPEG encoder shown in Figure 22.1.

Of concern if the subscriber records a program in a PVR, is the possibility of exchanging the program with someone else, either by transmitting the program on to the other person or by physically removing the hard drive from the host and putting it in another host or in a computer. For this reason, the content recorded on the hard drive must also be encrypted, unless it is marked for unlimited recording. As of this writing, an encryption standard has not been specified for this interface. The host manufacturer chooses one.

22.8 Middleware

A modern STT or host is as much software as hardware. In the beginning, STTs were limited to the manufacturer's software. Several proprietary sets of so-called middleware were developed; then standards began to emerge for certain levels of software. In particular, it is important to allow downloaded programs that will run on any STT to which they are downloaded so that an operator or subscriber may select from any of several STTs on the market. This can be achieved by a standardized *middleware* solution. The CableLabs/SCTE-standardized middleware is known as *OpenCable applications platform* (OCAP). As of this writing, two OCAP specifications exist, the OCAP 1 and 2 profiles.[30] OCAP is based on an earlier European initiative called *Multimedia Home Platform*, MHP.[31] There are a lot of similarities between OCAP, MHP, and DASE, the ATSC's *DTV Applications Software Environment*.[32] As of this writing, there are efforts to harmonize DASE, used in North American digital off-air broadcast, and OCAP, used in cable TV.

OCAP 1 closely follows MHP 1.0.2, but there are differences, some of which we will mention in this description.

22.8.1 Layers of Software

You may think of at least three layers of software running on the microprocessor(s) in a host. The lowest layer is the operating system, just as a desktop computer has an operating system. It is specific to the hardware being used and is typically purchased from a third-party vendor by the host manufacturer for the processor(s) chosen to be used. (This is consistent with the way OCAP is explained in the standards, but note that in some cases there is a layer of software below the operating system, call the *hardware abstraction layer* (HAL). The HAL allows the operating system to be platform independent.) The highest layer of software is the application software, which may, for example, provide an EPG in the form that the operator wishes to display it. Many other applications software packages are possible, from communications to games to messaging to VOD and others. Usually, application software is downloaded to a particular host based on what the user purchases and what the operator wants to provide. Application software is usually downloaded from a central server and may change from time to time.

The problem is that the operator needs to be able to download one software package for one application and expect it to run the same way on all hosts and STTs in the system. This must obtain even if there are different STTs and hosts made by many different manufacturers at different times, using different processors and operating systems, and having different capabilities. If a capability is supported by the hardware in a box, then the same software should enable it in all boxes. Making this happen is the role of *middleware*. The host manufacturer is responsible for providing middleware for his products that conforms to the appropriate OCAP specification, 1.0 or 2.0, as specified by the cable operator. It is possible to use a host that supports 1.0 but not 2.0. Each specification details the services to be supported and how relevant data is communicated to the host.

22.8.2 OCAP

OCAP 1 is an application interface that includes the *application program interfaces* (APIs) necessary to implement certain specified applications. It is intended to define a minimum software capability for OCAP-certified hosts sold at retail. Software providers may assume the existence of this level of capability when developing host applications. API is the set of routines, protocols, and tools for building software applications. A good API makes it easier to develop a program by providing all the building blocks. An application programmer puts the blocks together.[33] Thus, OCAP provides the programmer with all the tools needed to put together a certain application, which may then be downloaded to all STTs

and hosts, without regard to who made the device. The performance will meet some minimum standard of consistency on all devices.

OCAP 1 is based on a Java virtual machine model, a popular standard for the so-called *execution engine*. That is, you write Java programs to run on the host. Related terms you will encounter include *Hypertext Markup Language* (HTML), the authoring language used in creating Web pages. *Hypertext Transfer Protocol* (HTTP) defines how messages are formatted and transmitted to Web browsers and what actions Web servers and browsers should take in response to various commands. HTML is not a part of OCAP 1 (it is part of OCAP 2), but it is possible to use Java to write a Web display application if desired. Java can leverage the graphics capabilities supported by the hardware.

All applications in MHP are *bound*. That is, they are tied to the channel to which the host is tuned. If the viewer tunes off the channel, then the application stops unless it is also associated with the newly tuned channel. On the other hand, OCAP recognizes both bound and *unbound* applications. An unbound application continues to run even when the channel is changed.

Besides bound and unbound applications, and unlike MHP, OCAP allows native applications. Any function that can be programmed in Java (meaning just about any function) may be programmed and downloaded to an OCAP host. However, unlike bound and unbound applications, native applications are not guaranteed to run on any host. Native applications are written for a particular host.

Because it will not always be the case that the cable operator can control what applications are downloaded to a host, OCAP includes a *monitor application*, which decides what other applications are allowed to run on the host. This way, the operator has some control of what the subscriber does with his host and what the host might do to the network. An application can be downloaded on the FAT, the high-bandwidth channels on which video is downloaded. But the FAT channel is often passed through the headend without the operator's making any changes, opening up the system for other applications forwarded by the program supplier.

Figure 22.17 illustrates the relationship between OCAP 1 and other software and hardware in the host.[34] At the lowest layer is the OCAP host device hardware, what we traditionally think of as the host or STT. This is the collection of hardware on which the software runs. The lowest layer of software is the operating system/middleware layer. This includes a lot of traditional operating system functions, such as scheduling different activities, handling interrupts (such as when the subscriber presses a key), software to handle different interfaces, and memory management.

The second layer of the software is the execution engine, whose functionality and upward-bound interfaces are specified in OCAP 1. It provides prescribed interfaces with the application software above, through OCAP APIs. The upward-bound interfaces must be device independent. That is, any host, implemented with any software operating system, must respond in the same way when an application issues some command or supplies some piece of data.

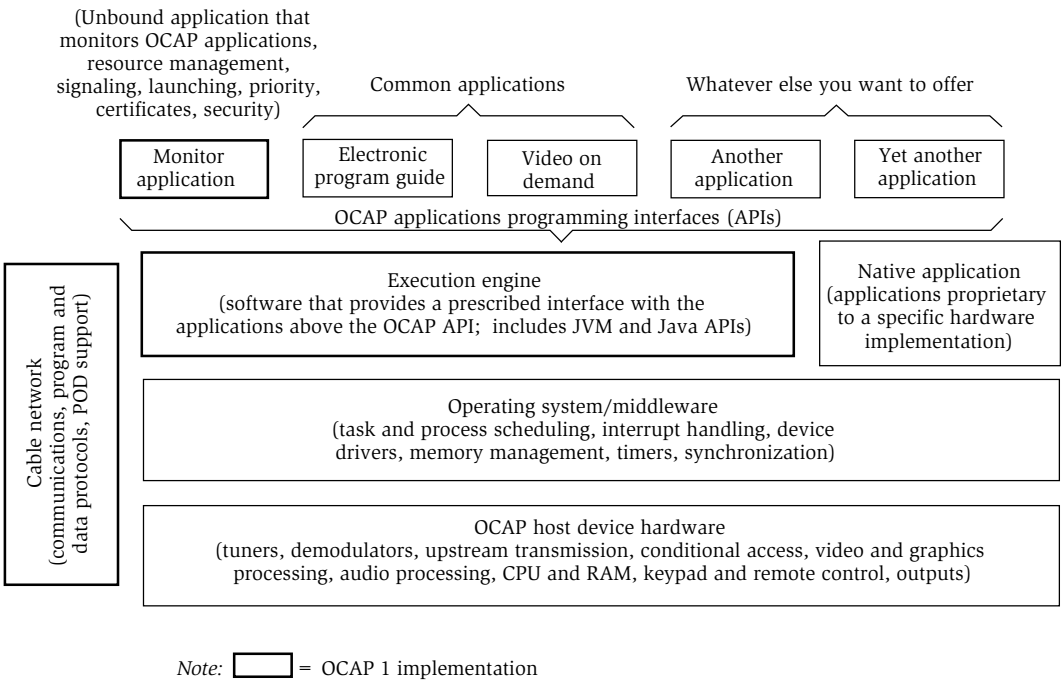

(Unbound application that monitors OCAP applications, resource management, signaling, launching, priority, certificates, security)
Common applications
Whatever else you want to offer
Monitor application
Electronic program guide
Video on demand
Another application
Yet another application
OCAP applications programming interfaces (APIs)
Cable network (communications, program and data protocols, POD support)
Execution engine (software that provides a prescribed interface with the applications above the OCAP API; includes JVM and Java APIs)
Native application (applications proprietary to a specific hardware implementation)
Operating system/middleware (task and process scheduling, interrupt handling, device drivers, memory management, timers, synchronization)
OCAP host device hardware (tuners, demodulators, upstream transmission, conditional access, video and graphics processing, audio processing, CPU and RAM, keypad and remote control, outputs)
Note: = OCAP 1 implementation

Figure 22.17 OCAP 1 software architecture.

Above the execution engine are the various applications, which might come from the host vendor but which might also come from other vendors. Common applications include EPG and VOD, but many other applications are possible. Some examples are listed later. The monitor application is included in OCAP 1. It controls what other applications may reside on the host, and it controls resource management between the applications and other functions that involve one or more applications. This is typically operator specific, because each operator will want to make his or her own decision about what runs on the hosts in the system.

The cable network is generally defined in SCTE 40, formerly DVS/313. It defines the levels and characteristics of the input and return signals interfacing on the left of Figure 22.1.

Native applications are any applications that run on a specific host and are not conformant to OCAP. Such applications may be written when the execution speed of an OCAP application is not fast enough (some games fall into this category). If an operator wants to enable such applications on several different hosts, he or she may have to write the application several times, once for each different host (assuming they are not compatible). In contrast, an OCAP application need be written only once, and it will run on all OCAP 1–compliant platforms.

OCAP 1 Representative Applications

A number of representative applications are presented, which OCAP 1-compliant systems are required to support. Typically these applications are written on top of OCAP 1 (using its services). This is not an exhaustive list — other applications can also be supported. There is no requirement that all of these applications be present on a particular host, but it is required that the box support software that implements these applications as appropriate. An OCAP 1–compliant box will be capable of having application software downloaded to support these services, but it is not a given that, just because a box is OCAP compliant, these applications exist on it. They exist when client software is downloaded. Also, every client might not have the hardware required to support every application listed.

Electronic program guide: Electronic program guide (EPG) software must be capable of being written using OCAP APIs, but OCAP does not specify a particular EPG. This is up to the software chosen by the operator. EPGs typically offer a grid showing programs available on the system at the present time and for some length of time into the future. The user can move through the grid, with current location being highlighted. If he or she presses a "SELECT" or "OK" key when on a particular program listing, the host tunes to that program. If the program is not on yet, the EPG may set the host to tune to that channel when the program does come on, or it may present a reminder to the viewer that the program is starting. The EPG may also be used for purchasing pay services such as pay-per-view (PPV) programs, which are presented at preset times and billed as an extra service. It can also be used for purchasing video on demand (VOD), a similar service except that the subscriber can order it to start when he or she wants it and can exercise VCR-like control (stop, play, fast-forward, rewind).

Watching TV: Though this might not be thought of as a software function, there are certain aspects of watching TV that might be controlled by applications written on middleware, and OCAP must support them. These applications support such functions as selecting standard- or high-definition viewing and, if in high definition, the format (aspect ratio, scan format, and resolution). The subscriber would want to control whether audio is delivered as mono, stereo, or 5.1-channel surround (based of course on what formats are transmitted). He or she might also want to select an alternate language or to select closed captioning information to be overlaid on the picture. Picture-in-picture, time shifting, emergency alert, and content advisory ("V-chip") are other services related to watching TV. Software to implement these services must be supported by OCAP.

Pay-per-view: Pay-per-view (PPV) is a traditional way of ordering pay programming. The subscriber contacts the system operator in advance to arrange for purchase of a particular event. The operator then downloads

authorization to the host. At the time of the event, the EPG may provide a reminder to the subscriber. Related services for delivery of pay programming include Impulse Pay-Per-View (IPPV), video on demand (VOD), and near video on demand (NVOD). These are described in Section 21.4. Additional services may be provided using the EPG capability of the host.

Email: Email clients are supported by OCAP. This would be an example of an unbound application, because it is not tied to any one broadcast channel. The operator is free to download an email client of his or her choice to an OCAP-compliant host.

Chat/conferencing: A chat or conferencing application would most likely be another unbound application, available no matter to what channel the subscriber was tuned. It might appear as a floating box over the video, in which the viewer and others could send instant messages to one or more other participants to discuss a show or anything else people can think of to chat about.

IP telephony: OCAP 1 specifically lists IP telephony support, though the host may not have the interfaces required for telephone service. IP telephony is covered in Chapter 6.

Games: OCAP 1 provides certain APIs required for game playing. These include windowed video and playing sound effects and animation. OCAP does not provide support for multiplayer networked games, but this capability can be provided in native application software. Some games may require faster response than provided under OCAP 1, in which case the game client may include native-code interfaces to support certain hardware.

Music/radio: OCAP provides for decoding audio without decoding associated video to permit you to transmit digital audio broadcasts. If you want to download audio in any other format, this could be done by downloading a software decoder. An audio decoding application would most likely be bound, because when the subscriber tunes away from the channel, the audio would stop.

E-commerce: These applications are usually broken into shopping and banking applications. Shopping benefits from high-quality full-screen graphics and flashy transitions and other effects to allow a merchant to present wares in the most pleasing way. Shopping also involves exchange of credit card information, so security is of importance. OCAP does not support a secure connection, leaving this to the application software. The operator must address privacy concerns. Banking services don't require the same high-quality video, but they do require extremely good security and user identification. Again, the application software must provide these services.

Minimum Functionality

OCAP also specifies certain minimal functionality that any host must support. Any OCAP host must support background, video, and graphics resolution of at

least 480×640 pixels, which is roughly equivalent to the monochrome resolution of an NTSC signal with a 4:3 aspect ratio ($4/3 \times 480 = 640$). This resolution must be supported for both 4:3 and 16:9 aspect ratios.

The minimum host must support certain decoder format conversions (DFCs). That is, when the signal arrives in one display format, the host must support conversion to another format supported by the display device. For example, if a picture arrives in a wide-screen (16×9) format and the display cannot handle wide screen, the host must be able to convert to a conventional 4:3 format. There are several ways this is to be done. A central portion of the picture may be sent to the display, or a letterbox format (black bands above and below the picture) may be sent. A number of specific processing options are required, as specified in the MHP documentation.[35] In addition, and different from MHP, an OCAP 1-compliant host must support a JMF (Java Media Framework) player for presenting subtitling or closed captioning information.

There is a rather lengthy list of remote control commands that are to be accepted. Support for both keyboards and remote control transmitters is recommended. A certain minimum amount of memory and the recommended architecture of that memory are included in the OCAP 1 specification. STTs and other hosts usually are running even when the user perceives power to be off. This is a standby mode, and OCAP recommends that the middleware recognize the difference between "standby" and "on" modes, because this is important to some applications.

22.8.3 OCAP 2

OCAP 2 is a second, related specification that builds on OCAP 1. Besides the functionality built into OCAP 1, OCAP 2 adds content formats based on Web technologies, such as XML, DOM, and ECMAScript. It is possible to provide complete functionality in an OCAP 1 device, but OCAP 2 opens additional options for providing that functionality.

Extensible Markup Language (XML) was developed by the World Wide Web Consortium,[36] an organization that develops technologies to expand the potential of the Web. XML allows designers to create their own customized tags, enabling the definition, transmission, validation, and interpretation of data between applications and between organizations.

The document object model (DOM) is the specification for how objects in a Web page are represented. The DOM defines what attributes are associated with each object and how the objects and attributes can be manipulated.

The European Computer Manufacturers Association (ECMA) has created a standard (ECMA-335) that allows applications to be written in a variety of high-level programming languages and executed in different system environments.[37]

22.9 Summary

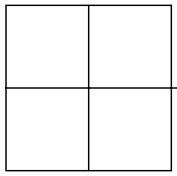
In this chapter we have presented a view of modern techniques used for digital video reception in the home. As this is written, many of the concepts and standards are new and have not been completely implemented in commercial equipment. The standard now is so-called *thin client* STTs, having limited functionality within themselves and depending on the headend for many functions. *Thick client* STTs, which include more processing power, have been used. The standards we have summarized can cover both. These standards are intended to support a wide range of functionality, to provide the cable operator with a rich choice of models through which to sell advanced services to subscribers. Of great importance is satisfying video content owners that their content will not be pirated. The standards anticipate the availability of devices besides STTs. For example, it is possible to subsume security needs into TV sets by using POD modules provided by the operator. Time will tell how the market for advanced services progresses.

In the next chapter, we discuss interfaces between consumer and cable TV equipment.

Endnotes

1. Cable Television Laboratories, *OpenCable™ Host Device Core Functional Requirements*, OC-SP-HOST-CFR-I11-021126, November 26, 2002. As of this writing, the standard is available at http://www.opencable.com.
2. www.scte.org.
3. Charles E. Spurgeon, *Ethernet: The Definitive Guide*. Boolder, CO: Cablelubs 2000.
4. Charles Poynton, *HDTV Video Short Course*, tutorial presented at the 2002 IEEE Consumer Electronics Society's International Conference on Consumer Electronics.
5. http://www.ddwg.org.
6. http://www.webopedia.com.
7. Another good reference is Luke Chang and Joe Goodart, *Digital Visual Interface*, May 2000. Available at www.dell.com — search for DVI.
8. Available from http://global.ihs.com.
9. Mark Eyer, EIA/CEA 775-A: The IEEE 1394–Based Digital Interface for DTV Receivers, *IEEE Consumer Electronics Society Newsletter*, August 2002, No. 3.
10. IEEE 1394-1995, *IEEE Standard for a High Performance Serial Bus*, available from www.ieee.org.
11. IEEE 1394a-2000, *IEEE Standard for a High Performance Serial Bus*, Amendment to IEEE 1394-1995, available from www.ieee.org.

12. Adam Kunzman and Alan Wetzel, 1394 High-Performance Serial Bus: The Digital Interface for ATV, *IEEE Transactions on Consumer Electronics*, Vol. 41, No. 3, August 1995.
13. Information for this section was provided by Jeffrey Riedmiller of Dolby Laboratories.
14. One asymmetric algorithm, RSA, is described in Justin Junkus, *Digipoints*, Vol. 2. San Francisco: SCTE, 2000, p. 169.
15. Bruce Schneier, *Applied Cryptography*. New York: Wiley, 1996.
16. Larry L. Peterson and Bruce S. Davie, *Computer Networks* San Francisco: Morgan Kaufmann, 1996.
17. Stephen Thomas, *SSL and TLS Essentials*. New York: Wiley, 2000, chap. 2.
18. http://www.digital-cp.com.
19. http://www.dtcp.com.
20. Bill Pearson, *Digital Transmission Content Protection* (slides), Intel Corp. 1999. Available at http://www.dtcp.com. Look for the DTCP tutorial.
21. Bruce Schneier, op. cit., section 16.2.
22. Bruce Schneier, op. cit., section 22.1.
23. Bruce Schneier, op. cit., chap. 12.
24. http://www.digital-cp.com.
25. The cryptographic schemes, primitives, and encoding methods are described in IEEE P1363.
26. DTLA, *Digital Transmission Content Protection White Paper*. Available at http://www.dtcp.com.
27. SCTE, *ANSI/SCTE 41 2001, POD Copy Protection System*. Available at www.scte.org.
28. Justin Junkus, *DigiPoints*, Vol. 2. Exton, PA: Society of Cable Telecommunications Engineers, 2000, p. 166.
29. Michael Adams and Tony Wasilewski, Multiple Conditional Access Systems, *Communications Technology*, November 1997.
30. See http://www.opencable.com. At some point these specifications may be transferred to the SCTE Web site.
31. See http://www.mhp.org.
32. See http://www.atsc.org.
33. http://www.webopedia.com.
34. OpenCable™ Application Platform Specification OC-SP-OCAP1.0-I04-021028.
35. ETSI TS 101 812 V1.2.1 (2002–06).
36. http://www.w3.org.
37. http://www.webopedia.com.



Chapter 23

Consumer Electronics Interface

23.1 Introduction

Cable services are enjoyed via the subscriber's consumer electronics products, the most important of which are television receivers and videocassette recorders. Stereo and surround-sound systems are also important to home theater installations. Coordinating the use of all the features of these products with cable services causes complications. Add-on devices such as video games, Web appliances, digital video disc (DVD) players, personal computers, home security and control systems, and other consumer electronics products add still more challenges. The cable–consumer electronics interface issues are complex, and satisfactory solutions have been elusive even with extensive committee work and cooperation between the industries. This chapter will survey the issues and some of the attempts to alleviate the problem.

There are two principal aspects to the issues concerning the interface between consumer electronics products and cable systems. One aspect is the way in which this interface affects the usage of cable services and of consumer electronics products and their features. The other aspect involves matters of law and politics. Both aspects will be considered, but separately.

Five major constituents have interests in these issues: (1) the subscriber/consumer, (2) the service provider, (3) the service provider's equipment manufacturer, (4) the consumer electronics and related industries, and (5) government agencies creating laws and rules related to these issues.

23.2 The Nature of the Interface

23.2.1 Cable Services

For purposes of this discussion, cable services can be categorized as (1) basic, (2) expanded basic, (3) pay, (4) pay per view, and (5) other.

Basic services are not scrambled as a matter of law. This is a requirement of the 1992 Cable Act, which we discuss later. Signal delivery is accomplished

merely by connection — usually by a technician physically connecting the drop cable to the cable plant. In a few cases, a switched tap is used to connect the subscriber under electronic control without the need for a technician to visit the tap location. Signal denial for those who no longer subscribe is done by disconnecting the subscriber from the cable plant. The subscriber who takes only the basic service and has a consumer electronics product that can tune all the channels of interest experiences the same amount of complexity as found when connected to an antenna.[1]

A converter with appropriate tuning capability is used in situations where (1) the subscriber's consumer electronics product does not tune all the channels of interest, (2) the product suffers from inadequate tuner performance, or (3) the product does not include remote control and the subscriber desires that feature. A converter is an extended capability tuner with a channel 3 or 4 output. It usually also includes a remote control. Converters were once commonly leased by cable operators. That is much less common today because converters are rarely needed since few very old receivers are still in service. Converters are available at retail.

Expanded basic services involve additional channels and an additional price. Signal denial for nonsubscribers is usually accomplished by traps or by scrambling. Alternative signal denial possibilities include interdiction. Interdiction has not found wide acceptance. All these technologies are described in the previous two chapters. When traps or interdiction is used, the subscriber experiences the same amount of complexity as found when the TV and VCR are merely connected to an antenna in the usual way recommended by the TV and VCR manufacturers. This assumes that the consumer electronics products tune all the channels of interest; otherwise, a converter is required.

Pay services, also called premium services, are sold by the month on a channel-by-channel basis. Pay services feature movies and special events and require a substantial additional fee. Analog signal denial for nonsubscribers is usually accomplished with traps for services with stable acceptance, that is, that have subscribers who keep their subscriptions for extended periods. There are two major types of traps: negative and positive. Negative traps allow subscribers to the premium service to receive it "in the clear" without any signal processing. With this technique, basic subscribers who do not take popular pay services have the negative trap installed in their connection to the cable system. The negative trap filters out the channel that is not taken. Popular services are those where more than half of basic subscribers take the service. Positive traps are installed at premium service subscriber locations when less than half of the basic subscribers take the service. A factor in making the choice between positive traps and negative traps is the need to minimize the number of units required. Alternative signal denial methods include interdiction and broadband descrambling. Conventional descrambling is used for services with higher rates of turnover or where additional security is desired. When traps, interdiction, or broadband descrambling is used, the subscriber experiences the same amount

of complexity as when the TV and VCR are connected to an antenna (assuming the consumer electronics products tune all the channels of interest).

Pay-per-view services are sold one at a time and almost always utilize an addressable descrambler. In the past, disposable traps that timed-out after the pay-per-view event were used, but with only moderate success.

Other services are generally taken by more sophisticated subscribers who have little or no difficulty with the interface between their equipment and the cable system.

23.2.2 Relevant Cable Set-Top Terminal Features

Cable set top terminals have a couple of important features that have been designed specifically to help with the interface between cable systems and consumer electronics products.[2]

The *auto bypass switch* is a feature that bypasses the entire cable spectrum to the TV (or VCR) when the set top terminal is turned off. This is an option that either is built into set top terminals or is available as an add-on. It is intended to allow the TV's (or VCR's) remote control to operate the TV's (or VCR's) tuner when the set top terminal is not used for scrambled channels. Though the auto bypass switch was available before the 1992 Cable Act's Consumer Electronics Interface amendment, the act motivated wider availability of the feature. Cable operators are required to make these switches available to subscribers but are allowed to charge for their costs. This feature has had little consumer demand.

The *dual descrambler* feature puts two tuners and descramblers in one package so that subscribers are not limited to unscrambled channels for some of the modes of operation of consumer electronics products. Cable operators are required to make these units available to subscribers and are allowed to charge for their costs. This feature also has had very little consumer demand.

The cable industry and independent entrepreneurs have developed a number of switching appliances to facilitate configuration of the interconnections for various purposes. Other devices intended to ameliorate these difficulties have been offered for sale. None of them has found long-lasting marketplace acceptance.

23.2.3 Relevant Consumer Electronics Characteristics and Features

TVs and VCRs are the primary consumer electronics equipment affected by cable service. These products come with a wide variety of features. The most important features for these issues relate to the remote control and the tuner. The remote control is used to turn the device on and off, change channels, and operate other features. Another feature of the VCR is its timer, which is used to commence and terminate recording and to change channels at appropriate times to capture programs of interest. Other TV and VCR features of interest include parental control, last channel tuning, sleep timers, and favorite channel access.[3]

Premium TV receivers include a feature called *picture-in-picture*, (PIP), which allows a small picture to be inserted into the large picture. Most often, the signal for the second picture comes from another source such as the tuner in a VCR or a television camera that might be placed in a baby's room or at the door to monitor visitors. Often the two signals may be interchanged with the press of a button. Another variation of the PIP feature includes the scanning of the channels with the capture of a set of miniature still frames providing a glimpse of what is on the various channels.

Most modern television receivers and VCRs have a feature, called *auto channel setup*, that automatically searches for available channels when the unit is first plugged in (and after every extensive power outage). The TV or VCR takes several minutes to search all the channels it can tune and loads into memory those that contain programming. Many (but not all) of these receivers will skip over most scrambled channels. When the viewer uses the channel up or down button, these channels will be skipped.

The *antenna and cable/converter switch* feature (usually operated by remote control) switches between two or more inputs to the TV. One input is intended to be from the converter and the other directly from the cable (or an antenna). The intention here is to allow the TV's tuner to do the work when nonscrambled channels are watched and just use the converter/descrambler's tuner when scrambled channels are watched. This feature is similar in function and purpose to the auto bypass switch option in converter/descramblers.

The *TV/VCR switch* feature on a VCR determines whether the entire cable spectrum is presented to the TV or just the channel tuned by the VCR or the tape playback.

23.2.4 Consumer Frustration

The two principal aspects of consumer frustration are the installation and the utilization of consumer electronics products with cable services. Both of these frustrations come from the variety of services and features found in these products.

23.3 Connecting Cable, TVs, VCRs, and Set-Top Terminals

In 1987, the NCTA Engineering Committee formed a Subcommittee on Consumer Interconnection and developed a set of 27 diagrams indicating the most likely methods of connecting a TV, VCR, and one or more set-top terminals along with an assortment of switches and signal splitters. Many of these diagrams are also found in consumer's manuals that come with new TVs and VCRs. All the interconnection schemes have advantages and disadvantages. None solves all of the problems. Reviewing some of these diagrams results in a better understanding of the complexity of the problem.[4]

When the original diagrams were developed, most TVs had 300-ohm balanced RF inputs, whereas today they have 75-ohm coaxial inputs. In the 300-ohm balanced input, you had to use a *balanced-to-unbalanced* transformer (balun) with a 4:1 impedance step-down ratio. This is not necessary with 75-ohm coaxial inputs, but we have retained the balun in Figure 23.1 since it is in the original document.

Where a VCR is shown, a *personal video recorder*, PVR, may be substituted. A PVR has an internal hard drive and saves programs on that rather than on a tape. Operation of a PVR is somewhat more problematical than that of a VCR because a PVR has added intelligence to allow it to select programs for later viewing, based on the subscriber's past actions. Having to impose an STT for tuning disrupts operation of this feature. Some STTs include PVR functions, so this becomes a nonissue in those cases.

Where digital services are to be received, an STT is required unless the TV has a cable interface to receive 64- and 256-QAM broadcasts directly without an STT. If so, then the STT may be omitted for nonscrambled digital programs. If the TV has a POD module interface as described later and in Chapter 22, it may be used without an STT in all cases. As of this writing, such TVs are not available. If the PVR has a POD module, it can be used on all channels without the STT. Again, such devices are not available as of this writing.

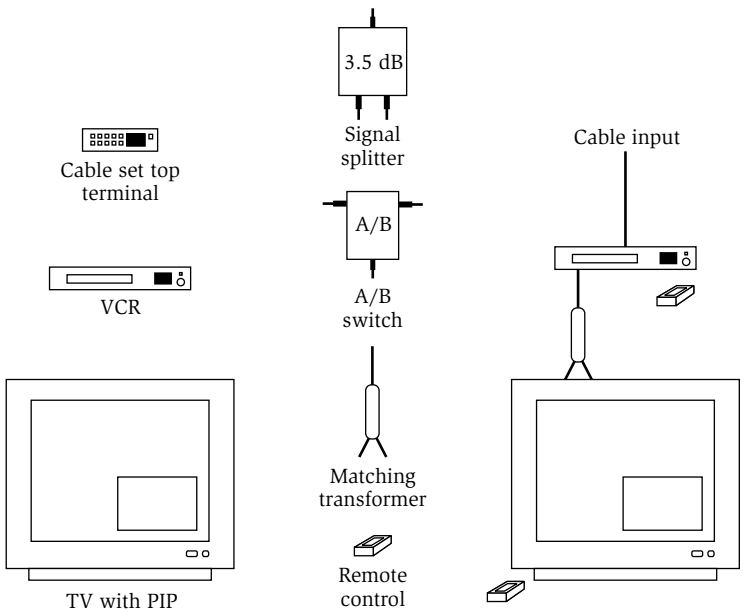


Figure 23.1 In-home equipment and simplest connection without set-top terminal.

Network-based PVRs may be implemented by the cable operator if desired, in which case broadcasts are saved on a hard drive at the headend and played out to subscribers as a VOD program. These would need an STT with appropriate software to control the network-based PVR.

23.3.1 Basic Cable and Trapped Premium Service Connection

Figure 23.1 is illustration 12 of the NCTA connection diagram set. The cable is connected to the VCR, which is connected to the TV. This diagram is for the basic-service-only subscriber and the subscriber who has premium services controlled by traps. No set-top terminal is used. Since less than half of cable subscribers have set-top terminals and since most subscribers connect more than one TV or VCR to cable, the majority of cable connections use this simple configuration.

Advantages

1. Least expensive.
2. TV and VCR operate as with antenna connections.
3. Can consecutively record different unscrambled channels automatically.
4. Can watch one unscrambled channel while recording a different unscrambled channel.
5. PIP works with unscrambled channels when VCR's baseband output is connected to TV's baseband input.

Disadvantages

1. No scrambled services available for either watching or recording.
2. No interactive services available.
3. Have to remember to put TV to channel 3 when watching recorded tapes.
4. Need to use two remote controls or a "universal" remote control.
5. TV's parental control is inoperative when using VCR's tuner.
6. Consumer's tuner subject to overload and damage from lightning strikes on cable.
7. Signal leakage possible from inadequately shielded tuners.

23.3.2 A Simple Start

Figure 23.2 is illustration 1 of the NCTA connection diagram set. The cable is connected to the set-top terminal, which is connected to the VCR, which is connected to the TV. This is by far the most popular connection method

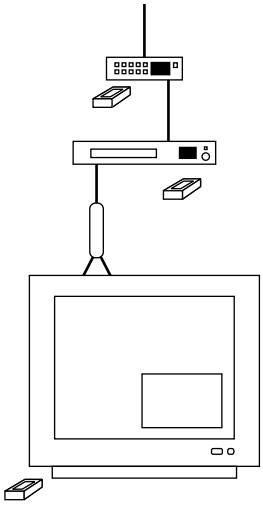


Figure 23.2 Most common cable connection with set-top terminal.

involving a set-top terminal because it is the least expensive and simplest to implement and use.

Advantages

1. Inexpensive and easy to install.
2. Easy to understand.
3. Can record scrambled channels.

Disadvantages

1. Can't watch one channel and record a different channel.
2. Can't consecutively tape different channels automatically because VCR tuner does not do channel selection.
3. Neither the TV tuner nor the VCR tuner has any control.
   (a) TV's parental control, favorite channels, and last channel recall features are inoperative. The TV is just tuned to channel 3 (or 4).
   (b) VCR's timer, parental control, favorite channels, and last channel recall features are inoperative. The VCR is just tuned to channel 3 (or 4).
4. Need to use three remote controls or a universal remote control to turn on the three devices and put the TV and VCR to channel 3 (or 4).
5. Makes the TV's picture-in-picture feature useless.

Special Case

1. If the converter/descrambler has an auto bypass switch, then when the converter is turned off, the entire cable spectrum passes through to the VCR, and the situation is exactly the same as in Figure 23.1, which is the case where no set-top terminal is used. When again using a scrambled channel, the subscriber must remember to put the VCR to channel 3 (or 4). If the set-top terminal has the channel-mapping feature and the TV or VCR does not, the same programming may appear on two different channel numbers. One channel number will be found on the set-top terminal and the other on the TV or VCR. This may be confusing.

23.3.3 Two Set-Top Terminal Case

Figure 23.3 is illustration 2 of the NCTA connection diagram set. The cable signal is split and connected to two set-top terminals. The left set-top terminal is connected to a VCR, which is connected to one input of an A/B switch, which is connected to the TV. The right set-top terminal is connected to the other input of the A/B switch.

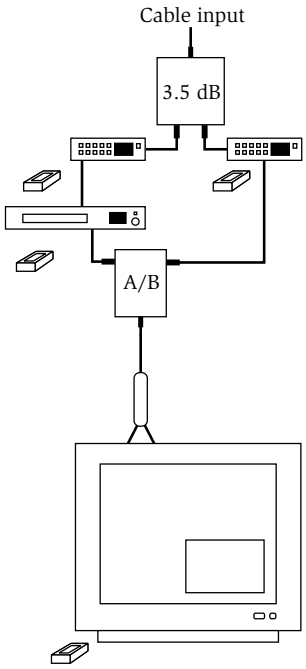


Figure 23.3 Connection with two set-top terminals.

This is one of the least popular installations because of its expense and complexity. It would be used by "techies" who want the most flexible system and can both visualize how it works and handle multiple remotes with ease. It will be very frustrating to others. The dual descrambler feature could be used to simplify both the installation and operation of this connection. Onscreen displays could further assist the subscriber. Despite the flexibility of this connection and the potential to make it relatively easy to use with onscreen displays, its usage is very limited.

Advantages

1. Most flexible.
2. Can watch any authorized scrambled or unscrambled program.
3. Can record any (other or same) authorized scrambled or unscrambled program.
4. PIP can be used with both channels if the TV has baseband inputs that are connected to the VCR's baseband outputs.
5. Useful in systems that have both analog scrambling and digital television.

Disadvantages

1. Can't consecutively tape different channels automatically because VCR tuner does not do channel selection.
2. Expensive hardware and monthly fees.
3. Complex to install, many cables, many opportunities for signal leakage.
4. Complex to operate; need to use at least three remote controls.
5. Neither the TV tuner nor the VCR tuner has any control.
   (a) TV's parental control, favorite channels, and last channel recall features are inoperative. The TV is just tuned to channel 3 (or 4).
   (b) VCR's timer, parental control, favorite channels, and last channel recall features are inoperative. The VCR is just tuned to channel 3 (or 4).
6. Need to use three of four remote controls or a universal remote control.
7. Most brands of set-top terminals will operate simultaneously from one remote control. Consequently, the subscriber may have to operate one terminal as manual. If the set-top terminal supports two control codes, four remote controls may be used with this configuration.
8. If both analog scrambled and digital channels are being used, may need to interchange set-top terminals depending on which is to be recorded and which is to be watched.

9. Switch at TV is usually manual, requiring getting out of the chair unless the switch is built into the TV and remote controllable.
10. Splitter cuts signal in half and may introduce noticeable noise.

Special Cases

1. If the left converter/descrambler has an auto bypass switch, then when the converter is turned off, the entire cable spectrum passes through to the VCR, and the situation is exactly the same as in Figure 23.1.
2. If the right converter/descrambler has an auto bypass switch, then when the converter is turned off, the entire cable spectrum passes through to the TV, and the situation is similar to that shown in Figure 23.1 but without the VCR.
3. If the set-top terminal has the channel-mapping feature and the TV or VCR does not, the same programming may appear on two different channel numbers. One channel number will be found on the set-top terminal and the other on the TV or VCR. This may be confusing.

23.3.4 Minimal Scrambled Channel Usage

Figure 23.4 is illustration 5 of the NCTA connection diagram set. The cable signal is split into two paths. The left path is connected to a set-top terminal, which is connected to an A/B switch, which is connected to a TV. The right path is connected to a VCR, which is connected to the other input of the A/B switch.

This connection could be used by someone who wishes to continue the use of the features of their TV and VCR as enjoyed before cable, and only occasionally watches, but never records, scrambled channels. This is especially convenient for someone with only one scrambled channel or who uses the converter/descrambler only for pay per view but doesn't record that programming. The right-hand path is identical to Figure 23.1. Alternatively, since the path with the converter can be used to watch either scrambled or unscrambled channels, the only purpose for the switch is to watch recorded tapes. That could also be accomplished with the VCR's baseband output and the TV's baseband input.

Advantages

1. Easy to use if most of the usage is for unscrambled channels.
2. Can watch any channel while recording any unscrambled channel.
3. Can consecutively record unscrambled channels automatically.
4. Especially easy if only one scrambled channel is taken.
5. PIP can be made to work as long as the second picture is unscrambled.

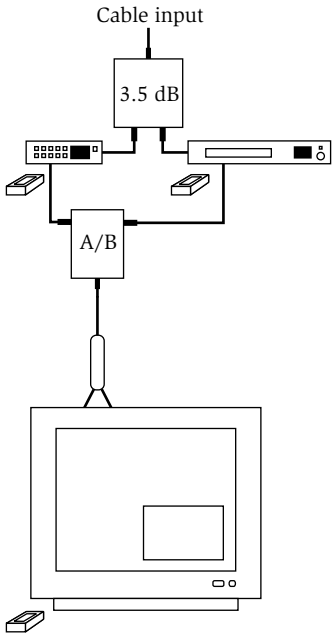


Figure 23.4 Connection for few scrambled channels.

Disadvantages

1. Can't record scrambled channels.
2. On the left side, TV's parental control, favorite channels, and last channel recall features are inoperative. The TV is just tuned to channel 3 (or 4).
3. Switch must be manually operated unless it is built into the TV.
4. Need to use three remote controls or a universal remote control.
5. Splitter cuts signal in half and may introduce noticeable noise.

Special Cases

1. If the converter/descrambler has an auto bypass switch, then when the converter is turned off, the entire cable spectrum passes through to the TV, and the situation is similar to that shown in Figure 23.1 but without the VCR.

2. If the set-top terminal has the channel-mapping feature and the TV or VCR does not, the same programming may appear on two different channel numbers. One channel number will be found on the set-top terminal and the other on the TV or VCR. This may be confusing.

23.3.5 Improved Simple Connection

Figure 23.5 is illustration 4 of the NCTA connection diagram set. This configuration inexpensively overcomes some of the difficulties of Figure 23.2 and especially makes sense if the TV has a built-in A/B switch.

Advantages

1. Inexpensive and easy to install.
2. Easy to understand.
3. Can record scrambled channels.
4. PIP feature can be used when one of the channels is always unscrambled and the VCR's baseband output is connected to TV's baseband input.

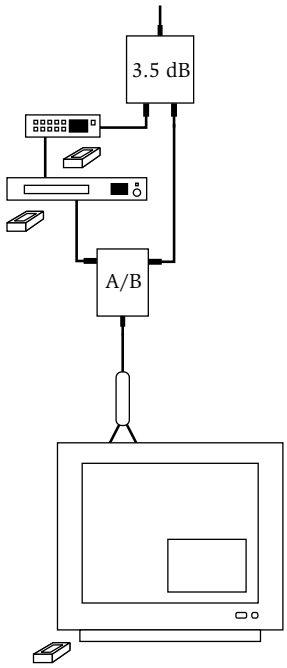


Figure 23.5 Modified most common connection.

Disadvantages

1. Can't consecutively tape different channels automatically because VCR tuner does not do channel selection.
2. Need to use three remote controls or a universal remote control.
3. On the left side, neither the TV tuner nor the VCR tuner has any control.
   (a) TV's parental control, favorite channels, and last channel recall features are inoperative. The TV is just tuned to channel 3 (or 4).
   (b) VCR's timer, parental control, favorite channels, and last channel recall features are inoperative. The VCR is just tuned to channel 3 (or 4).
4. Switch at TV is usually manual, requiring getting out of the chair unless the switch is built into the TV and remote controllable.
5. Splitter cuts signal in half and may introduce noticeable noise.

Special Cases

1. If the converter/descrambler has an auto bypass switch, then when the converter is turned off, the entire cable spectrum passes through to the VCR, and the situation on the left side is exactly the same as in Figure 23.1.
2. If the set top terminal has the channel-mapping feature and the TV or VCR does not, the same programming may appear on two different channel numbers. One channel number will be found on the set top terminal and the other on the TV or VCR. This may be confusing.

23.3.6 Problematic Connection Example

Figure 23.6 is illustration 7 of the NCTA connection diagram set. This figure is almost never used with modern set-top terminals because they usually require a separate frequency for receiving control signals. The VCR would block those signals, and the set-top terminal would turn itself off after a preset time. Also, this configuration would block upstream signals, making impulse pay-per-view (IPPV) and other two-way services impossible.

Advantages

1. Attempts similar functionality to Figure 23.4 but without the need for the A/B switch.
2. PIP can be used with the second picture unscrambled if the TV has baseband inputs connected to baseband outputs from the VCR.

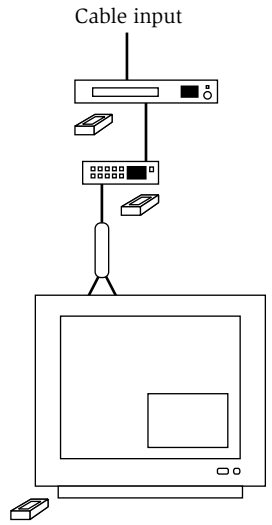


Figure 23.6 Connection with set-top terminal difficulties.

Disadvantages

1. Many set-top terminals can't be used in this configuration.
2. Impulse pay-per-view and other two-way services are impossible.
3. Can't record scrambled programs.
4. TV's parental control, favorite channels, and last channel recall features are inoperative. The TV is just tuned to channel 3 (or 4).
5. Need to use three remote channels or a universal remote control.
6. Converter/descrambler's parental control is inoperative when using VCR's channel 3 (or 4) output.

Special Cases

1. If the converter/descrambler has an auto bypass switch, then when the set top terminal is off, the entire output of the VCR passes through to the TV, and the situation is exactly the same as in Figure 23.1.
2. If the set-top terminal has the channel-mapping feature and the TV or VCR does not, the same programming may appear on two different channel numbers. One channel number will be found on the set-top terminal and the other on the TV or VCR. This may be confusing.

23.3.7 Improvements

One of the main disadvantages of the configurations shown in Figures 23.2, 23.3, and 23.5 is that the VCR is not able to consecutively record different programs except if the set-top terminal has an auto bypass switch and no scrambled programming is to be recorded.

This disadvantage can be overcome, at least in theory, with set-top terminals that include a programming function similar to that found in VCRs. This programming function can be set up to turn on and change channels at appropriate times. Usually, an onscreen menu is used for this purpose. The subscriber then uses the same principles to program the set-top terminal as are used to program the VCR, with the VCR programmed only to turn on and record the output channel of the set-top terminal at the appropriate time. In practice, this procedure meets with limited success by the occasional user since there are additional opportunities to make mistakes and thereby fail to record the desired programs. Nonetheless, the careful subscriber can master these techniques and successfully record different channels consecutively.

An alternative approach to overcoming the deficiency that inhibits the recording of consecutive programs on different channels is found in certain premium VCRs. These products have an infrared (IR) light emitter on a short wire. This device is sometimes called an IR blaster. The IR control is set up to use the IR control codes of the brand set-top terminal employed by the cable operator. The VCR can then change the channels of the set-top terminal and thereby record different channels, scrambled or unscrambled, consecutively. Since the set-top terminal is now controlled by the VCR, the set-top terminal's remote control is not necessary. Only two remote controls are needed by the subscriber, one for the TV and one for the VCR.

In the case of Figure 23.3, this type of VCR overcomes yet another difficulty. That is, if the two set-top terminals respond to the same IR codes, the VCR's IR controller can be placed in such a way that the set-top terminal used with the VCR is isolated from the IR signals used with the set-top terminal connected to the TV. Then only three remote controls are necessary, and both set-top terminals can be controlled remotely.

It is possible to use the same IR blaster technique with the TV controlling the set-top terminal (if the TV is so equipped). Also, the set-top terminal may use an IR blaster to control the VCR or TV. In that case, the subscriber programs the set-top terminal and uses its remote control.

An unfortunate complication with the IR blaster stems from the fact that the usual on–off control on consumer electronics products and set-top terminals is a toggle function. That is, there is not a separate on and a separate off control. Rather, if the unit is on, the control turns it off, and vice versa. This uncertainty over the current state is a source of error in setup and usage of these techniques.

There are numerous complications in these configurations. Most subscribers have difficulty visualizing what is happening with the multiple remote controls or the multiple modes of use of the universal remote control.[5]

23.4 Consumer Electronics Interface Issues

23.4.1 The Importance of Compatibility

The cable–consumer electronics interface establishes the manner in which a cable system, a consumer, and consumer electronics products function together to provide a useful result. Compatibility means the convenient use of the features of in-home equipment that are of interest to the subscriber. It also means a minimum of hurdles to subscribing to advanced cable services. From the perspective of the cable operator, compatibility is an issue of customer service and satisfaction. Effective compatibility can be a strong competitive advantage over other service providers. It can mean the difference between subscriber retention and the loss of subscribers to other media.

23.4.2 The Participants and Their Priorities

To better understand the consumer electronics interface issues, it is helpful to consider the problems from the perspectives of the various participants. These participants have differing goals and desires that are not always in harmony.

There are five major participants: (1) the subscriber/consumer, (2) the service provider, (3) the service provider's equipment manufacturer, (4) the consumer electronics industry, and (5) government entities making laws and rules.

The subscriber/consumer is the ultimate objective of all the efforts. If the subscriber/consumer is not happy and if there are alternative entertainment and information choices, the result will be a loss of business. Of course, there always are alternatives. The simplest alternative is to simply not subscribe to premium services and just use an off-air antenna with videotape and video disc rentals as supplementary video.

The next participant to be considered is the service provider who has a love–hate relationship with the in-home hardware. Following is the service provider's equipment manufacturer who has legitimate business concerns. The participant that is most anxious to become an active player in this arena is the consumer electronics industry. Lastly, the Congress has established laws that require the FCC to regulate these matters.

23.4.3 The Subscriber/Consumer

The subscriber/consumer participates in two ways. The subscriber/consumer buys equipment from the consumer electronics industry and subscribes to services that play on that equipment. The ultimate goal is the services. Additional programming choices make the TV and VCR even more useful.

The best reason for a large screen and a surround-sound system with multiple speakers and high-wattage amplifiers is the premium programming that comes

with pay services and pay-per-view events. Likewise, having a quality receiver and sound system makes movies and premium services more enjoyable. Given all this, it is clear that the cable and consumer electronics industries are highly motivated to work closely together to reinforce this symbiotic relationship.

The consumer wants good programming, good pictures and sound, cost-effectiveness, and simplicity. The subscriber does not care which medium makes the programming available. The consumer is frustrated with product features that are difficult to use and understand. The blinking "12:00" on the VCR has become emblematic of consumer technophobia — many consumers do not even know how to set the correct time. The maze of wires behind the TV and VCR are bad enough with just off-air reception, but they become beyond comprehension for many consumers when a premium service provider adds a set-top terminal of some sort.

23.4.4 The Service Provider

There are two elements to the service provider participant: the cable operator and the program supplier. Although this is an interdependent relationship, there are differences in their priorities.

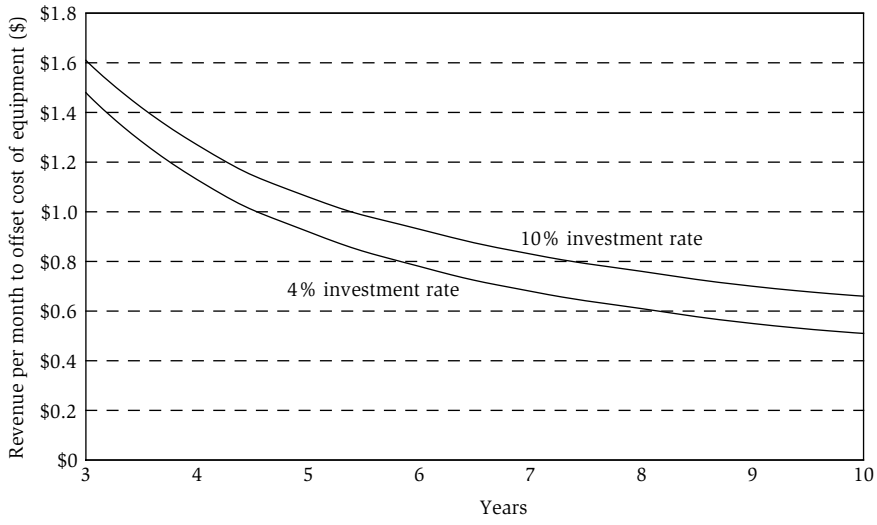
The service provider's love–hate relationship with the in-home terminal concentrates on its major disadvantage, the cost, and its major advantage, the facilitation of new services. The capital cost of the in-home terminal is a significant portion of the total capital investment required for service provision. This is especially burdensome with digital set-top terminals.

A rule of thumb used in the cable industry for a decade and a half is that for every $50 of in-home hardware, $1 a month is required just to cover the cost of the equipment. This rule of thumb is based on the assumption that the money to buy the set top terminal is borrowed from some capital source that requires repayment. The capital source may be internal or external, but capital is always a scarce commodity with multiple demands, each of which must justify its use. The capital must be repaid over the period during which the set-top terminal remains functional, and the interest must meet corporate hurdle rates. Over the years that set-top terminals have been used, the interest rate has varied over a wide range. Using the usual financial equations that are available in spreadsheet programs, Figure 23.7(a) shows a table of the amounts that must be paid back monthly to cover a $50 loan for various interest rates over a number of years. Figure 23.7(b) graphs this data with the number of years across the top, the dollars of outflow on the vertical axis, and a separate curve for each percentage value.

In the graph, the curves for the various percentages cluster closely. Clearly, the number of years over which the loan must be paid back is the more important parameter. Going from 3 years at 5% interest to 3 years at 10% interest increases the monthly cost by only 7%. Going from 3 years at 5% to 4 years at 5% reduces the monthly outlay by 23%. Long life of the equipment is

Years	Interest rate in percentage						
	4	5	6	7	8	9	10
3	$(1.48)	$(1.50)	$(1.52)	$(1.54)	$(1.57)	$(1.59)	$(1.61)
4	$(1.13)	$(1.15)	$(1.17)	$(1.20)	$(1.22)	$(1.24)	$(1.27)
5	$(0.92)	$(0.94)	$(0.97)	$(0.99)	$(1.01)	$(1.04)	$(1.06)
6	$(0.78)	$(0.81)	$(0.83)	$(0.85)	$(0.88)	$(0.90)	$(0.93)
7	$(0.68)	$(0.71)	$(0.73)	$(0.75)	$(0.78)	$(0.80)	$(0.83)
8	$(0.61)	$(0.63)	$(0.66)	$(0.68)	$(0.71)	$(0.73)	$(0.76)
9	$(0.55)	$(0.58)	$(0.60)	$(0.63)	$(0.65)	$(0.68)	$(0.70)
10	$(0.51)	$(0.53)	$(0.56)	$(0.58)	$(0.61)	$(0.63)	$(0.66)

(a)



(b)

Figure 23.7 Set-top costs. (a) Monthly cost of $50 loan at various rates paid back in multiple years. (b) Graph of data.

important to low monthly costs. Conversely, high-tech equipment that is soon obsolete is very expensive on a monthly basis. In addition to repaying the capital for the set-top terminal, there is a monthly maintenance cost.

The benefit of the in-home terminal to the service provider is that it facilitates services. The in-home terminal does this in two ways. First, it provides the electronics necessary for the service. For example, a Web appliance provides

the signaling, memory, and display electronics needed to browse the Web and display the results on the TV. Second, the in-home terminal is the instrument of conditional access. Unless access can be restricted, a price cannot be demanded. A wide variety of ways of protecting signals were discussed in the previous two chapters. Nearly all of them require some in-home device.

Experience shows that consumers respond to the ability to make impulse purchases of programming. The easier the mechanism of purchase, the more is bought. This capability also requires some sort of in-home equipment.

The in-home terminal is one of the fundamental determinants of the quality of the video and audio experienced by the subscriber. Technical specifications such as noise figure, linearity, and the tightness of the shielding are critical to good pictures and sound. The need for a quality terminal is paramount.

Lastly, there is a need to be able to experiment with new services. In the past, this has been done by the system operator providing the in-home hardware necessary for the marketing experiment. If the service proves compelling, additional equipment is procured, and the service is rolled out across the population. The cable industry is very good at copying success and applying it widely. If the experimental service is disappointing, the equipment is abandoned and forgotten. In neither case is the consumer required to purchase new hardware that might be later orphaned. This is just an extension of one of cable's major marketing advantages. The salesperson can offer a low-cost or even a free trial hookup to the consumer without asking for the purchase of any equipment. Consumers know that if they are dissatisfied with the service, they may discontinue at any time without getting stuck with hardware that has no other use. This ability to offer a risk-free trial is an important advantage.

The program supplier's principal concern is protection against theft. Though this is important to the cable operator, it is even more critical to the program suppliers. The framers of the U.S. Constitution realized the importance of protecting the rights of authors, artists, and other creative workers when they provided for the copyright. Protection encourages the creative contributor. Without the fruits of these contributors, there would be no cable industry and no consumer electronics industry. The consumer would have minimal interest in what little would be left. Although the public may view theft of programming as a "victimless crime" or one of little significance, this is a misperception. All suffer from the loss of incentive to produce quality programming. Legitimate consumers pay higher prices because those who steal do not contribute their fair share to covering the costs of operation. Fortunately, the 1992 Cable Act, the 1996 Telecommunications Act, and the actions of the FCC have recognized the importance of preventing signal theft.

The program supplier is also interested in minimizing the impact of the copying of programs on other means of distribution. Businesses in the rental and sale of videotapes and video discs are undermined by easy copying of programming delivered on cable.

23.4.5 The Service Provider's Equipment Manufacturer

The priorities of the service provider's equipment manufacturer are perhaps easiest to understand. Its goal is to profitably sell systems involving large quantities of in-home hardware and the headend equipment needed to supply signals. Since cable systems are individual closed systems of limited geographic area and since the in-home hardware has traditionally been leased, adjacent cable systems can utilize totally incompatible equipment and still well serve the TVs and VCRs owned by subscribers. The subscriber who moves from system to system is not inconvenienced by these incompatibilities since the system that serves the new residence will supply the necessary in-home equipment while the system serving the previous residence takes back its equipment.

A major advantage of this approach is that new technology is easily accommodated. When a cable system is ready to upgrade its hardware, it can purchase the best of the newest equipment available without concern about making the consumer's equipment obsolete. From the manufacturer's perspective, this keeps the market from becoming saturated. New services and new technologies can recapitalize the consumer's set top while bringing excitement to the service. Imprudently implemented consumer ownership of this equipment would interfere with this process of innovation.

23.4.6 The Consumer Electronics Industry

The consumer electronics industry has two main elements: the manufacture and the retail sale of products.

The nature of the consumer electronics business is one of high competitiveness. The manufacturer is interested in making more products and in selling upscale models. Almost 100 million residences receive television signals in one form or another in the United States. There are more than 250 million television receivers and 150 million VCRs in those homes. About 25 million new TVs and 15 million new VCRs are sold annually. The consumer electronics industry has the capacity to produce even more. Each year, new efficiencies make possible the production of more units from existing production lines. There simply is massive production overcapacity. The consequence of this is that it is impossible to maintain prices, much less raise them. Margins are tight. Only the full-featured products make money. The simple products now sell for something under $10 per inch of screen size. The cable set-top terminal has the potential to interfere with attractive features in the TV and VCR. This drives the consumer toward the profitless low end of the product line. Clearly, this is a survival problem for the consumer electronics industry.

The retailer does not always have the option of selling set-top terminals to cable operators. STT sales are traditionally large-scale industrial sales that are seen by the retailers as a bypass of their business. The retailer's desire is either to totally prevent the use of set-top terminals or to get into the business of selling

them. After a few years of attempting to outlaw set-top terminals, the retailers have come to realize that the age of interactive television, convergence, and advanced services requires set-top terminals. The emphasis is now on requiring these devices to be available at retail.

23.4.7 The Two Extremes

After the 1992 Cable Act, discussions between the cable and consumer electronics representatives began in earnest. Some on each side adopted extreme and simple-minded positions. The extremists in the consumer electronics industry said, "Why don't they just not scramble?" Whereas the extremists in the cable industry said, "Why don't they just make monitors?"

Both of these outlooks are life threatening to the industries concerned. Without scrambling, cable cannot protect signals and offer premium services. This would spell the end of the cable business. Without premium features, the monitor business would degenerate into a profitless commodity business that would cause chaos to the consumer electronics sector. Neither alternative is workable. Neither alternative considers the consumer. Something else is required.

23.4.8 The Origins of the Problem

The history of the situation provides some help in understanding its complexity. The first "cable-ready" TV was almost an accident. A couple of decades ago, television tuners were analog devices with mechanical rotary turrets. A multitude of contacts switched coils, changing the frequency tuned and the local oscillator frequency. Remote control was implemented with a motor driving the shaft of the tuner. The tuner contacts wore out or oxidized, and the tuner became the major source of service problems. The digitally controlled varactor tuner was an innovation that tuned the receiver with electrically variable capacitors (the varactors) instead of screwing a ferrite core into a coil of wire. The tuning voltage for the varactors was derived from a digital-to-analog converter driven by digital electronics. The voltage data was stored in semiconductor memory. The same digital electronics served to implement the remote control. The advantage to the consumer was a significant increase in reliability because of the elimination of moving parts and the vulnerable electrical contacts. However, the cost of TVs with electronic tuners was around $100 more than comparable units with mechanical tuners, making these advanced receivers difficult to sell.

An engineer recognized that memory comes in fixed sizes and that most applications do not use all the memory available on the chip. He proposed filling the memory chip with additional data that allowed the varactor tuner to tune some nonbroadcast frequencies used on cable. With the addition of a cable television F connector to accommodate connection to a cable, the product could be sold as "cable ready." The additional cost was near zero — just better

utilization of the memory that was already there. This was a feature consumers could understand. They were willing to pay more for "cable ready" because it promised to free them of the set-top terminal.

As time went on, subscribers noticed that though the TV gave them access to some cable channels, it did not give them access to all. A "horsepower race" developed where the number of channels tuned was part of the advertising. As more channels were added, a larger size memory was needed; this added a small increment of cost. Further increases in channel capacity required slightly more complex tuners. But this still was a minor cost increment.

23.4.9 Other "Cable-Ready" Issues

There are a number of important design features included in most cable set-top terminals that very few "cable-ready" TVs and VCRs incorporate. Most cable set-top terminals have a feature called *channel mapping*. Channel mapping decouples the channel display number from the frequency tuned. Thus, if there is a direct pickup problem in the cable system because of the presence of a strong broadcast signal, the broadcast channel can be placed at a different frequency to avoid this interference while allowing the subscriber to request it by its familiar channel number. The remote control command and the channel indicator all provide the familiar number while the tuner is "mapped" to another frequency. The set-top terminal has downloadable control commands to change the mapping as channel lineups change and to suit the needs of the cable system using the unit. A consequence of the absence of this important feature in "cable-ready" TVs and VCRs is the annoying need for multiple channel numbers in many local program listings. The familiar channel number is used for those subscribers using a cable operator-provided set-top terminal. An alternative channel number is used for "cable-ready" TVs and VCRs. This problem exists in the same home when some TVs have set-top terminals and others do not or when the auto bypass feature is used with a set-top terminal. Receivers built in compliance with the EIA-608A standard may include automatic channel mapping as an option.

A more important application of channel mapping is the manner in which scrambled channels are handled in a set-top terminal. When a subscriber using a set-top terminal attempts to tune a channel for which he or she is not authorized, the set-top terminal avoids that channel. Either it goes to another channel that has an appropriate message or it simply presents a blue (or blank) screen and muted audio. The alternative channel may include advertising for premium cable services. In that case, it is referred to as a *barker channel*. The term comes from the circus "barker" who hawked sideshows and attempted to attract an audience. If a blue screen is used, it may include a text or graphics message indicating how the subscriber could upgrade the subscription. In either case, the video and audio are kept from the receiver. This feature allows for practical video scrambling that has relaxed constraints. The video scrambling does not need to completely hide the video. There is no concern that "objectionable"

video might "bleed" through to offend sensitive viewers. Additionally, audio does not need to be scrambled since, in most cases, scrambling the video is sufficient to require a subscription. Children are protected from unintentional exposure to the audio of objectionable programming and do not accidentally see the "random lockup" of a scrambled video because of the set-top tuner's mapping to another channel.

The analog television audio signal is complex. It includes a stereo component, a second audio program (SAP), usually for monaural second language, and auxiliary signals. Since the human ear is very sensitive to audio distortion, scrambling the complex analog television audio signal in a manner that results in quality sound when descrambled is expensive and technically challenging. This expense would be unnecessary if channel mapping were universally used.

Set-top terminals have other important features as well. Parental control features are required by some franchises. These franchises often specify the manner in which the parental control feature must function and the level of security it must provide. "Cable-ready" TVs and VCRs manufactured since 2000 must have parental control (the "V-chip"). Additionally, set-top terminals often have an emergency alert feature that allows local government officials to alert citizens of hazardous conditions. "Cable-ready" TVs and VCRs do not include this capability, though it can be made available through use of a POD module.

23.5 Frustrated Product Features

Adding a set-top terminal either makes certain features and functions of TVs and VCRs redundant or interferes with their operation. In most situations, the feature or function is not blocked, but its use requires additional effort that the consumer finds inconvenient. It is this frustration that motivated the compatibility section of the 1992 Cable Act.

23.5.1 Remote Control

Remote controls were included in even early set-top terminals. The first remote controls were "wired." A long wire led to a separate unit positioned near the viewing chair and allowed channel changing without having to go to the TV. Cordless remotes based on ultrasonic tones, infrared light waves, or radio links were introduced later on. Consumers were able to add the remote control feature to existing non-remote-controlled TVs by utilizing the cable set-top terminal. This was seen as an important ancillary advantage to a cable subscription. The consumer electronics industry was not pleased with the slowing impact this had on replacement sales. However, a benefit accrued in that consumers were presold on the advantages of remote control. It became almost impossible to sell a new TV or VCR without remote control.

A problem arises when both the TV and the set-top terminal are remote controlled. The customer may feel he or she paid for a remote control feature on the TV that is not needed because the set-top terminal has the remote control. Worse yet, the consumer may need to juggle two remote controls. He or she may have to use the TV's remote to turn it on, put it to the output channel of the set-top terminal, and adjust volume. The consumer then uses the set-top terminal's remote to change channels. The set-top terminal often includes a switched convenience power outlet. This works well with older, non-remote-controlled TVs. But there are problems with remote-controlled TVs. These products usually have a toggle on–off push button, which means that simply powering the line cord does not result in the TV going on. Worse yet, if the TV has an automatic channel setup feature, it will go through the setup procedure every time it is repowered. Since this takes several minutes, the result is unacceptable. The net result is the juggling of two remote controls. Note that the same situation exists in a noncable application where the VCR is connected between the antenna and the TV.

23.5.2 Timers

When a set-top terminal is used with a VCR, further complications are encountered. Since the VCR timer controls only the VCR, it cannot turn the set-top terminal on and off, and it cannot change the terminal's channel. Early on, this was recognized, and set-top terminals with built-in timers have been available for well over a decade. The very same techniques used for programming the VCR are used for programming the set-top terminal. The consumer who can do one can do the other as well. Unfortunately, most consumers do this function so infrequently that they must relearn it every time they wish to record a program. The chances for error are great, and the success rate is low. This leads to frustration. The majority of consumers — cable and noncable — use VCRs for playback of prerecorded tapes and just avoid these headaches.

Some set-top terminals have the ability to set their clocks using signals coming from the cable headend. Thus, their time is always correct, even in spring and fall, when daylight savings time changes occur. Some newer VCRs implement this feature with signals from the broadcaster. Another common timer found in TVs is the sleep timer. Although a cable set-top terminal does not interfere with the TV's sleep timer operation, some set-top terminals have a sleep timer that makes one of the two timers redundant.

23.5.3 Picture-in-Picture

Picture-in-picture (PIP) is the insertion of a small picture within the larger picture. The purpose is to be able to monitor a second program while watching and listening to another program. If the action in the second program becomes compelling, the two images can switch places with the push of a button. Most TVs above the basic models include PIP.

There are a number of implementations of PIP. The most common embodiment has only one tuner, which is used for the main picture; the image for the second picture comes from an external device — usually a VCR's tuner. Other video source alternatives include a baby-monitor camera or a front-door camera. More expensive TV receivers have two tuners. An advanced form of PIP can be operated to scan through channels and put up a mosaic of small still pictures from a number of channels. As such, it is a kind of program guide that allows owners to scan the channels prior to making a selection.

In a TV with two built-in tuners, a problem arises if an internal splitter is used and the set-top terminal is placed ahead of the single antenna input. In that case, the output of the set-top terminal appears on both the large and small pictures. If on the other hand, two antenna inputs are provided, the use of the set-top terminal is noninterfering except when the customer wishes both pictures to come from scrambled channels. In that case, two descramblers are required.

The least common PIP implementation is the one that presents a mosaic of pictures from various channels. This system is interfered with since the program selector is in the set-top terminal and is not under the control of the TV receiver.

23.6 Early Attempts

There were a number of early attempts to solve or mitigate the compatibility problems between consumer electronics and cable systems. All proposed solutions were partial. The 80/20 rule applies. Eighty percent of the problem can be solved with 20% of the effort (and expense). However, solving the remaining 20% of the problem takes 80% of the effort.

23.6.1 Universal Remotes

Universal remote controls sought to alleviate the need for using two or more remote controls with the set-top terminal and TV or VCR. The universal remote control is configured for the brand TV, the brand VCR, and the brand set-top terminal. Simple universal remote controls have a hidden row of switches that are configured according to printed instructions that come with the device. An alternative approach is to enter appropriate setup codes from the printed instructions. Most models direct the channel change commands to the set top terminal, the volume commands to the TV, and the on–off commands to both units. VCR control can also be accommodated. Either a slide switch or a mode button governs which device is primarily controlled.

A problem arises with simple universal remote controls when a new model of TV, VCR, or set-top terminal becomes available or if a favorite feature of the TV or VCR is not included in the universal remote control. "Learning" universal

remote controls solve this problem by allowing the universal remote control to learn the signals emitted by the remote control that comes with the product. A disadvantage is the complexity; some consumers seem unable to master the instructions. Another type of learning universal remote control uses acoustic signals from the telephone to enter the data requested by the customer.

Advanced universal remote controls include *macros*. A macro is a collection of commands emitted in response to the pressing of just one button. For example, a common macro is the *advanced on* command. First, the set-top terminal is turned on, then the TV receiver is turned on, then the TV receives a command to tune to the output channel of the set-top terminal. It is important that all the devices "see" the complete IR emission sequence and respond appropriately, or the result may be disappointing.

Another model of advanced universal remote control includes a clock and a timer. The remote control can be positioned so that it simultaneously communicates with both the VCR and the set-top terminal. The universal remote control is programmed with the VCR control instructions. The remote turns on the VCR at the appropriate time, ensures that the VCR is tuned to the output channel of the set-top terminal, then tunes the set-top terminal to the desired channel, and finally starts the VCR recording. Consecutive recording of different channels is made possible.

A few models of two-way remote controls have a small screen on the remote control that displays signals received from the set-top terminal. Titles of songs, purchase order forms, interactive games, and other instructions can be viewed.

23.6.2 Electronic Program Guides

Electronic program guides (EPGs) have been introduced with features that alleviate some of the interface problems. The EPGs are available as stand-alone devices, built into cable set-top terminals, or built into TVs or VCRs. Several of these devices include one or more infrared light-emitting diodes (LEDs) on short wires that are meant to be attached near the remote control window of the VCR, the cable set-top terminal, or the TV. The device with the EPG controls the other devices to accomplish the desired result. If the EPG is built into the cable set-top terminal, the LED controls the VCR so that it turns on and off at the correct times and is tuned to the correct channel. If the EPG is built into the VCR, the LED controls the channel on the cable set-top terminal. Of course, none of these approaches can ensure that sufficient blank tape has been put into the VCR!

A related device, called VCR Plus, uses numeric codes published in printed guides (or conveyed electronically in a later version) to the VCR or to a form of universal remote control. The codes are decoded using a proprietary algorithm to derive the start and stop times and the channel number. Infrared emissions control the VCR and cable set-top terminal.

23.6.3 The TV Integrating Device

In the mid-1980s, the research department of a cable operator, American Television and Communications in Englewood, Colorado, began development of an approach the engineers called the TV integrating device (TVID). It was intended to be included in a cable set-top terminal and to use IR LEDs to control the TV and one or more VCRs. It had a built-in EPG and an on-screen display. The goal was to provide a simple menu that would allow subscribers to indicate their desires. The subscriber would indicate what he or she wanted recorded and what he or she wanted to watch. The TVID would take care of switching the descrambler into the correct part of the system and turn things on and off at the right time to implement the subscriber's desires. If a conflict arose, such as wanting to watch a scrambled channel while recording another scrambled channel, the subscriber would be advised of his or her choices and allowed to act upon them.

The TVID included a "mood guide," which gave the subscriber a way of indicating his or her mood. Those programs that were appropriate to his or her mood would then be displayed for selection to watch or record. The microprocessor could keep track of the programs watched and deduce the subscriber's interests. The subscriber would then be alerted to future programs that fit his or her previous viewing habits. The project reached the early prototype stage but was not completed because of other priorities. Some of the concepts, however, have become part of some PVRs.

23.7 Management of Expectations

Most of the subscriber's problems have solutions. Some solutions are more complex than others. The crux of the compatibility problem seems to be the expectation that the term "cable ready" brings. To consumers, this means that the TV or VCR can be connected to cable and all features and all services are readily available. When operation of features is complicated or unexpected results are obtained, frustration and anger set in. If instead, expectations were properly managed and the subscriber/consumer had a correct understanding of what was possible, the likelihood of disappointment would be reduced. The fundamental problem is the management of expectations.

23.7.1 Complexity Overload

A major contributor to the difficulties of the consumer electronics interface is the wide variety of choices available in features and functions, the way in which the features and functions are operated and controlled, and the diversity of services offered by cable operators. The number of combinations and permutations is almost infinite. Just the simple matter of entering channel numbers has several methods. Some products use an enter button. Others prescribe the way in which

the digits must be entered. For example, on some products, requesting channel 7 requires pressing 07. Others can require waiting for a fixed time after entering the command. The names for common functions are brand specific. The number of complex features boggles the mind. It may simply not be possible to preserve all the product functions and features while offering access to all the available cable services.

23.8 Legal Aspects of Cable–Consumer Electronics Compatibility

23.8.1 Compatibility Definitions

The word *compatibility* should mean that two things work well with each other. In the case of compatibility between cable service and consumer electronics, *compatible* should mean that the subscriber should be able to enjoy the cable services of interest and the features of the consumer electronics that he or she finds appealing. The subscriber should be able to conveniently do both.

Note that compatible is a situational definition. If a subscriber takes only basic (unscrambled) service from a cable system with a modest number of analog channels and is located far from strong interfering broadcast signals, he or she may enjoy all the functions of his or her "cable-ready" TV and VCR. If the subscriber then decides to add a scrambled service or moves near a broadcast transmitter, the situation changes significantly. He or she may lose compatibility between the TV or VCR and the cable system.

23.8.2 The 1992 Cable Act's Definition of Compatibility

Senator Patrick J. Leahy (Democrat from Vermont) drafted and introduced a bill in 1991 pertaining to the operation of cable systems and the design of consumer electronics products so as to improve compatibility. The bill did not pass. In the following year, it was introduced as Section 17 to the 1992 Cable Act.[6] Section 17 requires the Federal Communications Commission (FCC) to create rules to improve the compatibility between cable systems and consumer electronics products. The effect of this section of the law was to define compatibility as having three principal aspects:[7]

(i) to watch a program on one channel while simultaneously using a video cassette recorder to tape a program on another channel;

(ii) to use a video cassette recorder to tape two consecutive programs that appear on different channels; and

(iii) to use advanced television picture generation and display features

Although this definition of compatibility may not be complete, it does define a requirement for the FCC's rules.

The bulk of the 1992 Cable Act pertains to economic issues such as rate regulation, the rules for buying and selling systems, and the permissible business relationships between programmers and operators. A small, but very important part of the Cable Act concerned the interoperability of cable systems and consumer electronics products. Though the 1992 Cable Act gave the FCC a difficult assignment, it also laid down some guidelines. The FCC was told to consider the costs and benefits to consumers of its rules and the provision of effective protection against theft. But the ultimate goal was clear: the minimization of interference with certain special functions in consumer electronics devices when connected to cable systems.

The FCC was given some new authority to help it achieve these goals. The FCC's primary influence over the cable industry was granted in the following clause: "the Commission shall determine whether and, if so, under what circumstances to *permit* cable systems to *scramble or encrypt* signals or to *restrict* cable systems in the manner in which they encrypt or scramble signals."

The authority given the FCC over the consumer electronics industry is much more limited: "to specify the technical requirements with which a television receiver or video cassette recorder must comply in order to be sold as 'cable compatible' or 'cable ready.'"

To ensure that the goals of compatibility continued to be pursued in the future, the FCC was given continuing responsibility:

> The Commission shall periodically review and, if necessary, modify the regulations issued pursuant to this section in light of any actions taken in response to such regulations and to reflect improvements and changes in cable systems, television receivers, video cassette recorders, and similar technology.

Further requirements imposed on the operation of cable systems included obligations for the FCC to:

> promote the commercial availability . . . of converter boxes and of remote control devices to the extent technically and economically feasible, to offer subscribers the option of having all other channels delivered directly to the subscribers' television receivers or video cassette recorders without passing through the converter box.

The FCC was required to provide a report to Congress within a year of the passage of the act. This voluminous report was based on the extensive responses to the FCC's initial notice of inquiry on the subject. The FCC issued a notice of proposed rule making on November 10, 1993, and released its first Report & Order on May 4, 1994. In those documents, the FCC indicated its directions:

> [W]e clarify that the compatibility regulations being adopted herein apply only to cable systems and do not apply to . . . video dialtone services.

> Signal thieves [have been] notoriously successful . . . . We will not require cable systems to allow their subscribers to own descrambling equipment.
>
> [W]e expect that the Decoder Interface connector will be flexible enough to accommodate most, if not all, of the attributes of cable digital service.

The FCC realized that there is an extensive installed base of more than 11,000 cable systems, more than 250 million television receivers, and more than 150 million VCRs. In addition, approximately 100 million TV households purchase around 25 million color TVs and about 15 million VCRs each year. No law could bring compatibility to this situation in a short time. The FCC divided its approach into three periods: short-term measures, intermediate-term measures, and long-term measures.

In the short term, cable operators were required to avoid scrambling the basic services, which consisted of mostly the broadcast channels carried on cable. In addition, the cable operators were required to provide certain equipment such as switches to bypass the entire spectrum around the set-top terminal when the terminal is not required for descrambling, and timers in set-top terminals to change channels at specific times. Cable operators were given a year to make available multiple tuners and descramblers in set-top terminals. They were allowed to charge for the costs of this equipment and its installation. Additionally, the consumer electronics industry was given a year to discontinue the use of the terms "cable ready" and "cable compatible" or any similar terminology in the sale of TVs or VCRs that did not fully comply with the FCC's promised technical definition of "cable ready." A rule was included that would have required a label on devices that tuned cable channels but did not fully comply with the FCC's definition of "cable ready," notifying consumers that this device may under some circumstances fail to provide satisfactory performance when connected to cable. The consumer electronics industry managed to get this rule reversed and gained a relaxation of other rules in a subsequent petition for reconsideration submitted to the FCC. The FCC imposed a bizarre rule prohibiting the cable industry from changing infrared codes in each set top terminal installed in each home in order to protect the home owner's "investment" in purchased remote controls. This rule was all but impossible to follow because of the large number of different codes. It was removed in response to a subsequent petition for reconsideration. Also, the FCC recognized the aggressive work of those who would steal signals:

> [W]e are clarifying that the availability of set-top devices from retailers applies only to basic converters without descrambling or other access control function.
>
> [W]e do not believe it is desirable to require cable operators to make their security technology available for ownership by the public.

The FCC then separately considered longer-term measures centered around the decoder interface (described in the previous chapter) being developed by the

Joint Engineering Committee of the National Cable Television Association and the Electronic Industries Association: "We . . . emphasize . . . the Decoder Interface connector and associated component descrambler/decoders to be an important part of . . . equipment compatibility."

The FCC envisioned a digital world that would ultimately be based on standards and methods that avoided the problems of compatibility. The FCC intends to consider that environment separately in future actions.

23.8.3 The 1996 Telecommunications Act

The Communications Act of 1934 created the Federal Communications Commission and set the original public policy on communications. There have been multiple amendments to that act over the years, making it a quilt of patches. Additionally, technology had changed so dramatically that a complete rewrite became necessary. The result is the 1996 Telecommunications Act.[8] Although the act has a wide range of impact on cable television practice, only the portions pertaining to compatibility will be discussed here.

Probably the most important provision of the 1996 Telecommunications Act pertaining to compatibility is Section 304, "Competitive Availability of Navigation Devices," by Representative Thomas J. Bliley Jr., of Virginia:[9]

> (a) The Commission . . . in consultation with . . . industry standard-setting organizations . . . [shall] assure the commercial availability, to consumers of . . . services offered over multichannel video programming systems, of converter boxes, interactive communications equipment, and other equipment . . . from manufacturers, retailers, and other vendors not affiliated with any multichannel video programming distributor. Such regulations shall not prohibit any multichannel video programming distributor from also offering . . . equipment . . . if the system operator's charges . . . are separately stated and not subsidized.

The national headquarters of Circuit City stores is in Congressman Thomas Bliley's district, and his efforts to further the interests of an important constituent are clearly evident in this legislation.

The term *navigation devices* is a strange one. It has not been used in the cable or other related business except in reference to electronic program guides and related means of helping consumers find their way — navigate — through the wide variety of services available on systems with hundreds of channels. The law uses this term to mean any converter or set-top terminal intended to go between the service provider's terminal and the TV or VCR.

Several important points should be noted. Although the 1992 Cable Act was restricted to cable operators, the 1996 Telecommunications Act applies to any multichannel video programming distributor. Additionally, the services covered are not limited to video.

Limits were placed on what must be sold in order to protect the interests of programmers and system operators.

> (b) The Commission shall not prescribe regulations . . . which would jeopardize security of multichannel video programming and other services . . . or impede the legal rights of a provider of such services to prevent theft.

In order to accommodate experimentation with new services and technologies, the law provides for waivers allowing a period in which hardware can be under trial.

> (c) The Commission shall waive a regulation adopted under subsection (a) for a limited time upon an appropriate showing . . . that such waiver is necessary to assist the development . . . of a new or improved . . . service.

A number of significant hazards are brought about by this legislation. The most important hazard is that the other users of cable systems and of the over-the-air spectrum could suffer interference and degradation of service when consumers purchase their own navigation devices. Second, consumers are subject to confusion with these complex devices. Money may be wasted on features that are unusable or not needed. When technology changes, consumers will be stuck with orphaned hardware. In some cases, the ownership of equipment will create a reluctance to try new services that require additional hardware that partially duplicates what has already been purchased.

It is likely that many, if not most, subscribers will continue to lease equipment because of the many benefits that come from such an arrangement. The equipment is matched to the services provided. New equipment replaces old when technology evolves, service of equipment is included in the lease, and assistance on its use is readily available.

The other major portion of the 1996 Telecommunications Act pertaining to compatibility is Section 301(f), the "Eschoo Amendment" by California Congresswoman Anna Eschoo. This amendment modifies the compatibility rules of the 1992 Cable Act by putting restrictions on the FCC's actions. Specifically, it calls for mandating "a minimum degree of common design and operation, leaving all features, functions, protocols, and other product and service options for selection through open competition in the market." The goal of this amendment is to prevent the FCC's rules from picking winners and losers in a competitive environment where a variety of different approaches — some proprietary and others public — are vying for the same customer. Though that is a laudatory objective, it does make a difficult and complex task more challenging.

23.9 Standards

23.9.1 The Need for Standards

The cable industry has traditionally had little need for additional standards because it provides signals usable on the consumer's ordinary television or radio.

The goal has always been to allow the consumer to take cable service without having to purchase additional equipment. A major attraction to cable is the ability to try a subscription with little investment and the assurance that it is easy to drop the subscription if it is disappointing. So the most important standard is NTSC. When the cable system provides services beyond the ability of the consumer's TV, an adapter is provided to allow access to the additional services. The earliest adapters were just tuners that gave access to channels beyond the range of the TV receiver. Since a cable system is a closed universe unto itself, there is no need for interoperability with adjacent cable systems.

Conditional access and scrambling are deliberately not standardized. Standardization of these functions would give the signal thief a larger market to sell "pirate boxes." By having multiple types of conditional access and scrambling, the garage manufacturer of pirate boxes is frustrated. Since the service provider has ownership of the conditional access equipment, he or she can decide when the economic losses due to a breach of the security system are of sufficient magnitude to require replacing the equipment. When the service provider replaces the equipment, the consumers are not adversely affected.

The 1996 Telecommunications Act includes a section that requires the FCC to make rules that will regulate how multichannel video service providers will make in-home equipment available at retail. Now, for the first time, there is a need for standards. Consumers who purchase equipment for connection to cable need to be assured that it will work when they move to another cable system. Fortunately, the law and the FCC have recognized the problem of signal theft and have included a clarification that security systems do not have to be made available for subscriber ownership. There are three types of standards: mandatory, official voluntary, and de facto.

23.9.2 Mandatory Standards

Mandatory standards are those imposed by law or as a result of law. In the former case, the law itself contains the standards; in the latter, more common case, the law directs a government agency to do the difficult work and create rules. The FCC is the source of rules that govern the cable industry. There are two major sets of rules: (1) Part 78, the Community Antenna Relay Service (CARS) rules, which regulate the use of microwave links to convey signals across natural or man-made barriers or simply to transport the signals from one location to another where the geography in between does not have the cable service, and (2) Part 76, which governs the operation of the cable system itself. The items regulated include safety, signal leakage, and minimum performance requirements. These were first put into place in 1972 and then immediately stayed while further deliberations took place regarding the details. The performance rules cover frequency stability, carrier-to-noise ratio, allowable interference, in-band frequency response, hum content, adjacent channel response, cross modulation, must-carry and retransmission consent,

and signal leakage. The rules (with the exception of signal leakage) apply only to Class 1 signals, which are defined as signals originally broadcast. However, the rules cover only additional performance deficiencies. The cable operator is not responsible for problems in the signal as delivered to the cable headend. Class 2 signals are satellite-delivered programming, and Class 3 signals are locally originated. The rules were suspended by the 1984 Cable Act, which deregulated cable. At that time, they were converted into a recommended practice. Local governments were permitted to create their own rules, but those rules could not be more severe than the recommended practice. This was challenged by the League of Cities. A consequence of this was that a panel of representatives from the cable industry, the FCC, and the League of Cities negotiated a set of rules that has become the current FCC Part 76 in 1991.

23.9.3 Official Voluntary Standards

Voluntary standards do not have the force of law. The main motivation for compliance with the standard is the ability to use the fact of compliance in marketing efforts. Official voluntary standards have the sanction of a standards-setting organization. The primary standards-setting organization in the United States is the American National Standards Institute (ANSI). ANSI has very rigorous rules for the procedures involved in setting these standards. The rules are intended to ensure that participation is available to all interested parties and that due process is followed. ANSI certifies organizations as complying with these rules. Currently, the primary ANSI-certified cable organization is the Society Cable of Telecommunication Engineers (SCTE). Another organization that has worked to develop standards that govern both the cable industry and the consumer electronics industry is the Consumer Electronics Manufacturers Association (CEMA) sector of the Electronic Industries Association (EIA). Although CEMA is the official ANSI standards-setting organization, its efforts are under the auspices of the Joint Engineering Committee (JEC) of the EIA and the National Cable Television Association (NCTA) Engineering Committee.

The SCTE has about 100 standards under development at the time of this writing. Five of those standards are in a condition where they are suitable for submission to ANSI. One of those standards has been approved as ANSI/SCTE PS SP 400 1996 F-Port (Female Outdoor) Physical Dimensions. Next will be the 5/8–24 RF Outdoor Fitting Physical Dimensions standard. The Digital Video Transmission Standards for Cable Television is expected to be submitted before the end of 1998.

The JEC efforts began in the very early 1980s and have not been nearly as successful in completing standards. This lack of results is an indication of the difficulty involved in the interface between cable and consumer electronics. Its first standard, the ANSI/EIA-563 Decoder Interface Standard,[10,11] was in place for several years and then officially withdrawn. Another standard, EIA-542 Cable Television Channel Identification Plan,[12] lists channel numbers

and frequencies for three different channel plans: the standard plan related to the broadcast frequency plan, the harmonically related carrier plan, and the incrementally related carrier plan. Though this standard has been slow to make its way through the official process, it has been used by both the cable and the consumer electronics industry as a de facto standard for more than a decade. Both industries have found it a useful means of cooperation.

One of the most important JEC standards is the EIA-IS-105[13] Decoder Interface Standard described in the previous chapter. Other standards in the works include the National Renewable Security Standard (NRSS)[14] and the RF Interface Standard IS-23.[15] An effort to create a digital television standard has been transferred to the SCTE. The IS-23 standard governs such parameters as the physical connection to cable, the frequency response, the signal levels, and direct pickup by inadequately shielded components.

23.9.4 De Facto Standards

De facto standards are the result of marketplace forces where the cable industry has settled on a particular approach because it works well. The types of connectors, cables, amplifiers, and other equipment used in cable have been sorted out by the marketplace, and a limited set of models have evolved on a de facto basis. Some of these de facto standards are being converted into official voluntary standards by the SCTE. This approach has served the industry well and will continue to be in use.

23.10 OpenCable

Cable Television Laboratories, or CableLabs, has undertaken a program intended to provide options for alleviating many of the interface problems. The goals of the program, called OpenCable, are to provide sufficient standardization of features and functions so that a multivendor interoperable environment can be created. CableLabs will certify compliance with OpenCable. OpenCable functions are described in Chapter 22.

23.11 National Renewable Security System

The Joint Engineering Committee formed a subcommittee called the NRSS subcommittee to consider methods that could overcome a security breach. The motivation is to facilitate the sale of set top terminals or consumer electronics products while protecting the conditional access system by having it renewable in the event the system is defeated. The initial approach involved the Smart Card system, which was used for conditional access and for credit and debit cards. The Smart Card was intended to include a digital key that would decrypt the digital signal.

After some discussion, it was decided that the only safe approach is to have every transistor involved in the security function replaceable in a module that plugged into the set-top terminal. A wide variety of proposals were considered. One of the proposals used the Smart Card physical format but put a lot more electronics on it. This NRSS Type A card complied with the International Standards Organization (ISO) 7816 standard. Several limitations flow from this. There are only eight contact pads, forcing the MPEG signal to be transported serially. There is no mechanism for handling cable's out-of-band signaling. Eventually, the committee settled on an optional Type B approach based on the Personal Computer Manufacturers Communications Interface Adapter (PCMCIA) Type II standard card. This card has 68 pins, which is adequate to allow the set-top terminal to convey out-of-band signals received by a built-in modem.

The NRSS standard was designated the EIA-679. NRSS provided the basic idea for the POD module concept described in Chapter 22.

23.12 Summary

The interface between cable service and consumer electronics products affects the quality of service and the ease of use of a service. In severe cases, the interface prevents access to a service. This same problem interferes with the ease of use of some of the functions and features of consumer electronics products. These problems are complex because of the wide variety of services, functions, and features possible with modern methods.

Endnotes

1. Walter Ciciora, Consumer Electronics Friendly Cable, *1992 NCTA Convention Technical Papers*. NCTA, Washington, DC.
2. David Large, Options for Friendly Delivery of Cable Services. *CED*, April 1993, pp. 62–74, 83.
3. Walter Ciciora, The Realities of the Consumer Electronic Compatibility Interface Regulations, *1995 NCTA Convention Technical Papers*. NCTA, Washington, DC.
4. David Large, *Connecting Cable Systems to Subscribers' TVs and VCRs — Guidelines for the Cable Television Industry*. NCTA Engineering Committee, NCTA, Washington, DC, January 1987.
5. Walter Ciciora, Regulation and the Continuing Struggle for Consumer Electronic Compatibility, *1995 Montreux International Television Symposium Technical Record*. Montreux, Switzerland, June 1995, pp. 597–611.

6. Walter Ciciora, The 1992 Cable Act and the Consumer Electronics Interface, *1994 NCTA Convention Technical Papers*. NCTA, Washington, DC.
7. Cable Television Consumer Protection and Competition Act of 1992, House of Representatives Report 102–862, 102nd Congress, 2nd Session.
8. Telecommunications Act of 1996, House of Representatives Report 104–458, 104th Congress, 2nd Session.
9. Walter Ciciora, The Realities of the Retail Sale of "Navigation Devices," *1996 NCTA Convention Technical Papers*. NCTA, Washington, DC.
10. *EIA/IS-15-A, Standard Baseband (Audio/Video) Interface Between NTSC Television Receiving Devices and Peripheral Devices*, March 1989. Electronic Industries Association, Engineering Department, Arlington, VA.
11. *The EIA-653 Decoder Interface Standard*, 1990. Electronic Industries Association, Arlington, VA. Available from Global Engineering Documents, Englewood, CO, 800–854–7179.
12. *EIA-542 Cable Television Channel Identification Plan*, 1997. Electronic Industries Association, Engineering Department, Arlington, VA. Available from Global Engineering Documents, Englewood, CO, 800–854–7179.
13. *EIA/IS-105 Decoder Interface Standard*, 1997. Electronic Industries Association, Engineering Department, Arlington, VA. Available from Global Engineering Documents, Englewood, CO, 800–854–7179.
14. *EIA/IS-679 National Renewable Security Standard*, 1997. Electronic Industries Association, Engineering Department, Arlington, VA. Available from Global Engineering Documents, Englewood, CO, 800–854–7179.
15. *EIA/IS-23 RF Interface Specification for Television Receiving Devices and Cable Television Systems*, 1997. Electronic Industries Association, Engineering Department, Arlington, VA. Available from Global Engineering Documents, Englewood, CO, 800–854–7179.

Other References of Interest

David Large, Defining "Cable Ready" Consumer Equipment. *CED*, April 1993, pp. 84–90.

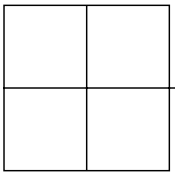
EIA-608-A Recommended Practice for Line 21 Data Service, 1994. Electronic Industries Association, Engineering Department, Arlington, VA. Available from Global Engineering Documents, Englewood, CO, 800–854–7179.

Several feature articles on the decoder interface. *Communications Technology*, August 1989.

The Consumer Interface Problem, A Status Report, CableLabs, Louisville, CO, February 1990.

Customer Premises Equipment Performance and Compatibility Testing, CableLabs, Louisville, CO, 1993.

Subjective Assessment of Cable Impairments on Television Picture Quality, CableLabs, Louisville, CO, 1992.



Chapter 24
Equipment Compatibility

24.1 Introduction

The previous chapter described the nature of the challenges regarding the interface between cable and consumer electronics equipment. That chapter concentrated on the way the equipment is connected and used, as well as on the impact on usability of features and services. The legal requirements and FCC rules pertaining to usage of the equipment were also covered.

This chapter will cover the minimum technical requirements for consumer-owned equipment that is to be connected to cable. There are two broad aspects to this issue: (1) the quality of the pictures and sound delivered to the subscriber, and (2) the potential for interference with other devices connected to cable and with other users of the radio spectrum. The first aspect is considered a marketplace issue to be determined by the purchaser of the equipment. But this aspect of compatibility unfortunately also affects the enjoyment of cable-delivered services and, to some extent, the likelihood that the consumer will subscribe to premium services. Often, the consumer purchases his or her equipment in an environment where its performance on cable is impossible to judge and factor into the purchase decision.

24.1.1 Broadcast Versus Broadband Receivers

The technical requirements of tuners for broadcast channels versus broadband cable channels are significantly different. It is possible to design an excellent broadcast tuner that performs poorly on cable channels, and vice versa. Designing a cost-effective tuner that performs well in both environments is a significant challenge.

The broadcast environment consists of channels with gaps. Local channels are almost never in adjacent frequencies. However, the viewer may wish to receive a weak signal with frequency between two strong local signals. The weak signal may be from a distant location. This is called *fringe reception* and was an important selling feature before cable became popular. The tuner must function well in that environment. Note that the lower channel numbers, the

VHF channels, are designated by the FCC so that the frequency of no one channel is a second harmonic of any other. This reduces the requirements on linearity in the tuner circuits.

Broadcast channels are not located on an "image" channel of another broadcast channel. The image channel problem will be described later in this chapter. Image interference problems are avoided in the broadcast environment. An inspection of the frequencies assigned to the VHF channels reveals that no channels are assigned so that they are images of valid channel frequencies. The higher channel numbers, the UHF channels, have an elaborate set of "taboos" in the FCC's rules that are intended to relax the design requirements of practical consumer electronics tuners. Finally, in the broadcast environment, there is no particular need to prevent direct pickup of the broadcast channel. Direct pickup simply adds to the signal brought in from the antenna and causes no harm. Also, since only broadcast channel frequencies are found in the tuner, there is no need to prevent their radiation from the tuner terminals or from the tuner itself.

The cable tuner is very different. The cable spectrum is fully loaded with no significant gaps up to the highest frequency. The signal levels are strong and relatively uniform across the frequency band. The challenge is in handling a lot of energy and in avoiding distortion and cross modulation of the signals.

A consequence of these differences is that a properly designed cable tuner will probably not perform well in the broadcast environment, and a properly designed broadcast tuner will probably not perform well in the cable environment.

Essentially, all cable converter tuners are double conversion tuners. That is, they apply the superheterodyne principle twice. They have two local oscillators and two intermediate frequency amplifiers and filters. This gives improved selectivity and blocks the image channel. Also, the first local oscillator frequency is usually above the highest channel tuned, and the second local oscillator signal is isolated from the cable by the first conversion stage. So backfeed of internally generated spurious signals is not a problem with cable converters. The use of a double conversion tuner in a "cable-ready" TV or VCR is extremely rare. Nearly all consumer electronics tuners are single conversion devices.

The deficiencies of "cable-ready" tuners will become more apparent as cable systems expand bandwidths beyond 450 MHz on to 750 MHz and even 1.0 GHz and beyond.

24.2 Levels of Interference

Four kinds of interference have been considered. In order of severity, they are (1) interference with the product's own performance, (2) interference with other products in the same home, (3) interference with another subscriber's reception, and (4) interference with other users of the electromagnetic spectrum.[1]

Interference with the product's own performance results in a dissatisfied customer for the product and probably for the services delivered. Unfortunately,

in many cases, the poor performance of the equipment can be confused with a problem with the service. This results in unnecessary and unproductive service calls, missed work while waiting for the service call, and the potential for finger-pointing between the cable service provider and the retailer of the consumer product. Too often the consumer gives up. The problem is not resolved, and the consumer is suspicious of both the consumer electronics product and the cable operator.

The next level of severity occurs when the product backfeeds interfering signals into the cable, causing problems with another TV or VCR in the same home. This is often a transitory problem occurring only when certain combinations of channels are tuned by the two devices. This is likely to be blamed on the cable operator. Yet one step worse is when another subscriber suffers interference from a product not even in the same home. This will almost always be blamed on the cable operator.

The most severe situation is when interference is caused to other users of the radio frequency spectrum, particularly safety-related users such as aircraft communications and navigation, fire, police, and emergency communications. This level of interference goes beyond merely reducing the entertainment value of cable and can become dangerous to others' well-being.

24.3 Direct Pickup Interference

One of the most troublesome problems is direct pickup (DPU) interference.[2] DPU in equipment connected to cable is caused by inadequate shielding of tuner and related circuits and loose connections that allow signal ingress on channel frequencies. A common DPU occurrence is the location of a receiver in the strong field of a transmitter whose signal is also carried on cable at or near its broadcast frequency. Since the propagation time in cable is more than the propagation time in free space, the over-the-air signal gets to the receive site first. A short while later, the cable signal arrives. The two mix in the receiver, and a deteriorated picture results. The consequences can range from a usable but degraded picture to a loss of the channel. The degradation can take the form of a ghost of the channel and/or beats between offset visual carriers. If the interference is relatively weak, only the sync pulse will be visible as a vertical bar in the picture. If the broadcast signal is not carried on the same frequency in the cable, the cable video will be different from the ingress signal. The result can be a moving background pattern. When the pattern is weak, the sync bar of the interferer wipes through the picture like a windshield wiper. When the interfering signal is strong, both pictures may be visible moving past each other. Depending on the phase of the ghost's visual carrier, the image may appear as a distorted color negative.

DPU is not limited to the simultaneous reception of a broadcast television signal over the air and through cable. In other parts of the spectrum, signal such as from amateur radio or pagers can make its way into the receiver and cause

problems. In these cases, the interference manifests itself as alternating dark and light bands that usually move through the picture.

It is important to appreciate that a receiver properly designed for broadcast reception may have no need for extensive shielding. It may perform well in the broadcast environment. When connected to cable, the result might be unacceptable. The consumer may then draw the wrong conclusion. He or she may deduce that since antenna operation yields good performance and the connection to cable has difficulties, there must be something wrong with the cable system. A technically sophisticated observer is not likely to fall prey to such a failure in logic. An intelligent, but nonexpert, consumer is likely to make such a mistake and complain to the cable operator.

24.3.1 Estimate of Exposure

CableLabs and the EIA funded research by consultants to study the problem. Stern Telecommunications Corporation of New York City did a study of the potential impact of DPU. The study included a literature search, measurement of representative recent "cable-ready" receivers, and a computer model of transmitter power levels at various distances from the transmitter.

The receiver testing was done at the Carl T. Jones Corporation in Springfield, Virginia, in July 1992. Five receivers were tested. The testing revealed that there is a range of susceptibility to DPU. This susceptibility is a complex phenomenon that depends on the orientation of the receiver relative to the signal's field, the way in which the coaxial cable and the power cord are oriented with respect to the signal's field, and the channel involved.

The computer modeling considered the 10 top broadcast-relevant geographic areas, called areas of dominant influence (ADIs). The ADI terminology is important in considering advertising reach and effectiveness. For the purposes of this study, the ADIs have well-documented data on the distribution and density of households. This geographic data is important to the estimates of exposure to DPU. The 10 ADIs include about 30% of U.S. households. The field strength contours were computed with a commercially available program by Communications Data Service, Inc., which is based on the FCC's propagation curves. The study considered only UHF channels up to channel 27 (visual carrier at 549.25 MHz) on the assumption that most cable systems are limited to 550 MHz. As such, the study is a little out of date, but the general trend is clear. Field strengths of 80 dBu (10 mV/m), 90 dBu (31.6 mV/m), 100 dBu (100 mV/m), and 120 dBu (1 V/m) were considered.

Data was presented for each of the 10 top ADIs and extrapolated to all U.S. households. Table 24.1 lists the percentage of all U.S. households that may be subject to at least the field strengths listed from stations operating up to 550 MHz.

This data shows that the previously used Canadian specification of 100 mV/m will leave a significant number of receivers vulnerable to DPU. Even at an

Table 24.1 Exposure to various levels of fields

Frequency Band	*80 dBu 10 mV/m*	*90 dBu 31.6 mV/m*	*100 dBu 100 mV/m*	*120 dBu 1 V/m*
VHF	61.5%	41.0%	19.9%	0.9%
VHF and UHF	65.4%	54.8%	40.8%	6.0%

immunity level set at 1 V/m, a modest number of subscribers will still either have to accept DPU or take a set top terminal with better shielding.

24.3.2 Estimate of Susceptibility

In December 1992, CableLabs solicited proposals to develop test methods for measuring DPU susceptibility and to measure a representative grouping of current products used in the home. The Carl T. Jones Corporation was hired for both of these tasks. Thirty-five television receivers, 8 VCRs, and 13 cable set-top terminals were tested for DPU susceptibility. The products were subjected to fields of increasing strength until a DPU 55 dB below the desired channel's sync tip was achieved. The 55-dB figure was confirmed to be reasonable by perception of interference studies. The discussion in Chapter 2 on the W curve and DPU indicated the 1993 tests generally confirmed the results of the 1986 CBS experiments. Across the three channels, four frequencies, two carrier-to-noise ratios, and two desired signal levels, the average value of the sine wave strength attenuation relative to the carrier strength to achieve "just perceptible" impairment levels was 56 dB.

Table 24.2 summarizes the results for the sample set of products. The mV/m figures required to breach the −55-dB threshold for slightly worse than "just perceptible" interference are listed for the best, worst, and median products. As can be seen, for the 1993 model products tested, a wide range of results were obtained.

An important difference between set top terminals supplied by cable operators and consumer electronics products is that the cable technician has several

Table 24.2 Performance of units tested

Susceptibility	*TV*	*VCR*	*Set-Top Terminal*
Best	193.4 mV/m	162.73 mV/m	732.33 mV/m
Median	51.57 mV/m	57.27 mV/m	292.73 mV/m
Worst	1.17 mV/m	29.9 mV/m	77.23 mV/m
Best/Worst	165	5.4	9.5
Sample Size	35	8	14

units with him or her and will keep trying them until the DPU problem is solved. Since the consumer purchases a TV or VCR and takes it home, the ability to pick the best from a collection is not available. Also, when the consumer makes a purchase decision, that selection will usually be in a different neighborhood from his or her residence. The retailer may not be using a cable feed for display purposes. Often, a laser disc is used as the source in order to provide the best possible pictures to encourage a purchase.

24.4 Receiver Performance

The Joint Engineering Committee (JEC) of the National Cable Television Association (NCTA) Engineering Committee and Consumer Electronics Manufacturers Association (CEMA) of the Electronic Industries Association (EIA) developed an RF Interface Specification for Television Receiving Devices and Cable Television Systems. This specification has the designation EIA-IS-23.[3] Eight parameters of customer premises equipment are important to this specification: (1) reradiation of cable signals, (2) local oscillator leakage and backfeed, (3) A/B switch isolation, (4) DPU backfeed, (5) VCR through-loss, (6) adjacent channel rejection, (7) image rejection, and (8) tuner overload distortions.

Carl T. Jones Corporation developed tests for these eight parameters. The test procedures were reviewed by the JEC in October 1993. Although no formal acceptance of the test procedures was solicited, comments were noted, and the tests adjusted where appropriate. There were two goals for this work: to develop test procedures for these parameters, and to determine the current state of the art.[4]

24.4.1 Reradiation of Cable Signals

Cable systems operate in an enclosed spectral environment. If this environment is sealed, the same frequencies that are used in the over-the-air spectrum for other purposes can be reused for cable channels. In fact, multiple enclosed cable systems can be simultaneously operated to further increase the number of signals delivered to subscribers. As long as the cable system remains sealed, no harm is done to other users of the over-the-air spectrum.

A potential source of signal leakage exists in any equipment connected to the cable system. This is because the entire cable spectrum is conveyed to the input terminals of the connected device. If leakage of these signals occurs because of inadequate shielding of the circuits, the potential for interference with other users of the spectrum exists. This is especially serious in the case of frequencies used for aircraft navigation and communications and for emergency messages. Cable systems occupy these frequencies in their closed systems and cause no harm as long as the integrity of the system remains intact.

Equipment connected to the cable system must also be designed to prevent signal leakage. If this equipment leaks signals, interference may result. Set top

terminals purchased by cable operators are specifically designed to prevent such leakage. The output of these set top terminals is just one television channel that is on a frequency not used locally for over-the-air service, usually channel 3 or 4. If leakage of this channel occurs from the TV or VCR connected to it, no harm is done since this is an unused local frequency. Consumer electronics products designed for broadcast applications do not have this requirement. When used for broadcast reception, they do not tune frequencies that are used for other purposes; they just tune the broadcast channels. Because of these differences, it is important to impose additional requirements on products connected to cable and prevent the connection of equipment that is not specifically designed to avoid leakage.

The FCC rules, Part 76.605 (a)12, limit the leakage or egress of RF energy. Cable systems must not emit energy that causes more than 15 μV per meter at a distance of 30 meters for frequencies less than 54 MHz or greater than 216 MHz. Between these frequencies, the limit is 20 μV per meter at 3 meters distance.

Of the units tested, only two television receivers failed.

24.4.2 Local Oscillator Leakage and Backfeed

Television receivers generate internal signals that are used to process the received spectrum. In addition, other internal signals are generated to display information on the screen and to operate functions and features of the product. Switch mode power supplies are also a source of interference. If these signals backfeed into the cable system, they can cause interference and degradation of performance in other TVs and VCRs connected to the cable system. They can add to the background noise, which makes two-way cable operation difficult. The other TVs and VCRs on the same signal splitter in the same residence are most at risk since there is a minimum of attenuation of the interfering signals. However, neighbors' receivers on the same cable system tap are also subject to this interference. Consequently, signal backfeed is carefully specified in equipment used by the service provider. Consumer equipment requires the same degree of care if problems are to be avoided.

The local oscillator in the single conversion tuners commonly used in TVs and VCRs operates at 45.75 MHz above the desired channel's visual carrier frequency. This frequency is in the frequency range of the cable system for most of the channels tuned. If this signal leaks back into the cable system, interference will result to other receivers according to the W curve discussed in Chapter 2. Fortunately, the splitters used by cable operators are required to have a minimum of 18-dB isolation between ports and usually have 20 dB or more. Thus if the local oscillator or other signal backfeed is no worse than 35 dB below the visual carrier, the 20 dB of the splitter will increase the signal attenuation to 55 dB below the visual carrier when it reaches the other receiver connected to the same splitter. At this level, it will be "just perceptible," but not annoying. Any further losses in the connecting cable will also help. Cable set-top terminals avoid this problem by using double conversion tuners, with the first local oscillator frequency higher than the cable spectrum tuned.

24.4.3 A/B Switch Isolation

A/B switches are part of the input of many receivers and VCRs, and allow remote-controlled selection of input sources. These sources include external antennas; cable signals; and the outputs of set top terminals, video games, and other playback devices. If insufficient isolation exists between ports of the switch, cross feed of signals will occur. The FCC regulations dated 1987 in Part 15.606 require a minimum of 80-dB isolation for frequencies between 54 and 216 MHz; 60 dB is required above that to 550 MHz. Since 1987, cable systems have gone beyond 550 MHz. Switches that have inadequate isolation at these higher frequencies will be a source of difficulty.

When an antenna is used, two kinds of problems may result. Antenna input signals can exceed +25 dBmV, whereas cable signals can be as low as 0 dBmV — lower if the subscriber does his or her own wiring and has split the signal excessively without appropriate amplification. In those situations, the isolation must be adequate to prevent interference. This explains why the cable industry uses switches that generally exceed 90 dB and often 100 dB at all used frequencies. The other problem arises when strong cable signals are not sufficiently isolated from a rooftop antenna. If cross feed of the cable signal occurred in many homes, the cumulative effect of efficient radiators all pointing in the same direction, that is, toward the broadcaster's transmitting tower, could yield unacceptable interference to aircraft traversing that path.

Although most of the devices tested had no problems with this specification, there were three instances that failed.

24.4.4 DPU Backfeed

The DPU backfeed issue is related to the DPU phenomenon but is not the same problem. A receiver that has a DPU problem may also backfeed such signals into the cable to cause difficulties with other devices. Again, there is a benefit from the 20-dB isolation between signal splitter ports. DPU backfeed of less than −35 dB will be at the perceptible but not annoying point.

Unfortunately, 37% of TVs had DPU backfeed above the perceptibility point on channel 6, 34% on channel 12, 26% on channel 78, and 17% on channel 59. Understandably, units with DPU problems had DPU backfeed problems as well.

24.4.5 VCR Through-Loss

Designing VCRs with adequate playback signal quality is a difficult technical challenge. A technique used in early VCRs to mitigate the problem was to unsymmetrically split the signal at the input terminals of the device, giving the larger share to the recording circuits and shortchanging the "through" side of the splitter. The through side of the splitter provides an attenuated version of the entire input spectrum to the RF output terminal of the VCR when neither playback nor the VCR tuner is used. The attenuation is due to the use of a splitter

without first amplifying the signal. This trick helped the noise performance of the recording circuits while providing a less disadvantageous comparison for the viewer when watching unrecorded materials.

As the last chapter discussed, the most common configuration of set-top terminal, VCR, and TV connects the cable system to the set-top terminal, which is connected to the VCR, which is connected to the TV. The TV suffers the additional through-loss of the VCR. This through-loss should be no greater than 5 dB. All eight tested VCRs passed this hurdle.

24.4.6 Adjacent Channel Rejection

The broadcast environment avoids adjacent channels. Cable utilizes as much of the spectrum as possible with nearly all channels having adjacent neighbors. An important issue for cable tuners is the ability to separate the desired channel from the upper adjacent visual carrier and the lower adjacent aural and chroma carriers. Cable practice is to operate the aural carriers some 17 dB below the visual carriers. This additional attenuation relaxes the difficulty somewhat. The criteria again is the 55 dB used to bring the signal to the not-annoying level.

The passband of the receiver is a combination of the IF response and the tuner response, which differs slightly from channel to channel and is under the influence of the automatic gain control (AGC) strategy. Consequently, devices may have isolated points of failure on some channels but not on others.

Of the devices tested, three television receivers and one VCR failed to sufficiently reject the lower adjacent color subcarrier. Two television receivers failed the aural rejection test on all channels, and two others had problems on just two channels. Of the eight VCRs, three had no adjacent aural problems, three others failed on the four channels tested, one failed on three channels, and one unit had difficulties with just one channel. All TVs and VCRs behaved adequately with the upper channel's visual carrier. Not surprisingly, cable set-top terminals passed all tests since they are specifically designed for such service.

24.4.7 Image Rejection

Image channels are a result of the superheterodyne receiver design, which uses a balanced mixer with two inputs followed by a filter for the purpose of separating out the desired signal from the spectrum at its input terminals. The inputs to the balanced mixer are the received spectrum and a local oscillator. The output of the mixer consists of the sum and differences of the input frequencies.

TV local oscillators for VHF are 45.75 MHz above the desired visual carrier. The difference between the local oscillator frequency and the desired frequency is the intermediate frequency (IF), which passes through the filter characteristic of the IF amplifier. The signal is then detected and becomes the source for the picture and sound selected by the viewer. The sum of the local oscillator and the desired channel is blocked by the filtering of the IF amplifier.

There is another difference frequency equal to 45.75 MHz. Signals that are 45.75 MHz above the local oscillator frequency will also have a difference of 45.75 MHz relative to the local oscillator frequency. This signal, if present at the input of the mixer, will be heterodyned to a frequency that will pass through the filter characteristic of the IF amplifier and provide interference to the picture and/or sound of the desired program. The visual carrier of the desired channel is 45.75 MHz below the local oscillator frequency. The image signal will be 45.75 MHz above the local oscillator frequency or 91.5 MHz above the visual carrier of the desired channel, about 15 6-MHz channels away. The visual carrier is the strongest part of the interfering channel. Visual carriers are spaced 6 MHz apart in cable practice. The visual carrier 15 channels above the desired carrier will be 90 MHz higher in frequency. When local oscillator frequency is subtracted from it, the visual carrier in the image region will be 1.5 MHz above the desired visual carrier and produce an interfering beat in the picture. The aural carrier of the channel that is 14 channels above the desired channel will also fall into the video bandwidth of the desired channel.

Since the tuner determines the rejection at the image frequency, the level of rejection varies with the channel. All receivers did well on channel 3, and only four had problems with channel 12. Half the receivers had interference on channel 53, and 67% failed on channel 74.

24.4.8 Tuner Overload Performance

Tuner problems and other deficiencies may result in impaired performance on the TV or VCR itself. In addition to the DPU problem previously described, nonlinearity in the receiver's circuits can give rise to the production of new frequencies not present in the original signal. The signals thus generated produce either diagonal bars in the picture or even a discernable image from another channel.

Tuner overload is a serious problem that will become more serious as cable bandwidth expands. Of the TVs, 66% had problems on at least one channel, 50% of VCRs suffered as well. Surprisingly, even 36% of the set-top terminals had troubles.

24.4.9 Rejection of Signals in the Upstream Band

Two-way services, once rare, are now commonly offered by cable operators. High-speed data service, two-way video set-top boxes, sometimes telephone interface units, and various types of status monitoring combined fill up a bandwidth whose upper limit has increased from 30 to about 42 MHz in recent years to meet the demand for additional upstream capacity. The result is that signals are generated at fairly high levels within customers' homes and transmitted toward the headend. Due to the finite isolation offered by taps and in-house splitters, some of that energy is directed toward the input terminals of television receiving devices.

The problems are twofold. First, the amplitudes may be sufficient to cause input stages to distort, producing intermodulation products that may fall within the normally tuned channels. Second, the upper frequencies fall within the common television intermediate frequency band of 41–47 MHz and may leak directly into sensitive amplifier stages if not sufficiently rejected by the input circuits.

The magnitude of the reflected signals can be estimated as follows. The power level of transmitted signals from standard cable modems are controlled by the headend and must be adjustable to as high as +58 dBmV to meet the DOCSIS specification (though some models can go higher). For reasons discussed in Chapter 16, higher levels are consistent with reduced average levels of ingress, and cable operators design their plants so that as many modems as possible are transmitting near their maximum capabilities. Using the assumption in Section 24.4.2 that the sum of cable loss and tap or splitter isolation will be at least 20 dB, the maximum reflected power at the antenna terminals of a television receiver in the same or a neighboring house can be as high as +38 dBm.

The presence of signals of that magnitude falling below 30 MHz should not be a problem because television receivers are designed to reject signals below that frequency with magnitudes as high as −7 dBm (+ 42 dBmV). At the other extreme, some receivers offer little protection for signals falling at or above 41 MHz, with levels as low as +10 dBmV causing visible interference in some models. IS-23 is silent on the issue of receiver protection above 30 MHz and also silent regarding maximum levels in that range that may be delivered by cable operators. Clearly, however, without some form of additional protection or avoidance of at least the frequencies above 41 MHz, many receivers will generate unacceptable levels of picture degradation when nearby two-way devices are transmitting near the upper end of the upstream band. The testing described earlier did not include this critical parameter.

24.5 Regulatory Status

The issue of consumer equipment compatibility was addressed in Section 17 of the 1992 Cable Act, which directed the Federal Communications Commission to create rules[5,6] "to specify the technical requirements with which a television receiver or video cassette recorder must comply in order to be sold as 'cable compatible' or 'cable ready.'"

In response, the FCC issued a Notice of Inquiry in January 1993 requesting input on a variety of consumer equipment interface issues, including the issue of "cable-ready" receivers.[7] Almost a full year later (December 1993), it issued a Notice of Proposed Rulemaking (NPRM) covering the same issues. This time, the key elements of draft EIA-23 were jointly submitted by the Electronic Industries Association and the NCTA,[8] supported by the CableLabs compatibility testing results (submitted by InterMedia Partners).

Finally, in May 1994 the FCC issued rules[9] defining the minimum performance characteristics of equipment marketed as "cable ready" or "cable compatible." These were amended in April 1996. While largely adopting some of the provisions in the interindustry agreement, the commission made several critical changes. First, they declined to adopt a shielding standard that would ensure that reradiation of cable system signals from connected receivers would be below the limits allowed cable systems. Second, they declined to control any aspects of the quality of devices with extended-range tuners specifically designed to tune cable channels, unless they were labeled specifically as "cable ready" or "cable compatible," even though some of the requested quality provisions existed purely to ensure the integrity of the system to which they were attached. As a practical matter, there has been no enforcement of even these weakened provisions.

24.6 Summary

There is a substantial difference in the technical requirements for good performance in the broadcast environment and in the cable environment. Receivers that perform well in one environment will likely have difficulties in the other. Designing a cost-effective product to do well with both types of signals is very difficult. We have described the difficulties and the consequence of failure. Testing reveals that much of the equipment available at the time of the tests did not make the grade in several aspects.

In summary, the relevant parameters for equipment connected to cable are at a minimum:

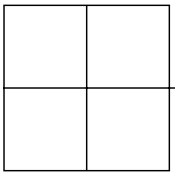
- Immunity to external fields of at least 1 V/m.
- No more reradiation of cable signals than is allowed by FCC leakage rules.
- At least 80-dB isolation between switch ports.
- DPU backfeed into the cable system at least 35 dB below the desired visual carrier.
- Through-switch attenuation of no more than 5 dB.
- Rejection of adjacent channel carriers to levels at least 55 dB below the desired visual carrier.
- Image signal rejection of at least 55 dB below the desired visual carrier.
- Tuner overload performance so that spurious signals are at least 55 dB below the desired visual carrier.

The challenge of making products that function well in both environments is a heavy one. There is no enforced set of standards for awarding the title "cable ready." There is no compliance testing service. Much of the consumer's

difficulty stems from not appreciating that extended tuning range television receivers are not necessarily fully compatible with properly designed and maintained cable television systems.

Endnotes

1. *The Consumer Interface Problem, A Status Report*, CableLabs, Louisville, CO, February 1990.
2. *Customer Premises Equipment Performance and Compatibility Testing*, CableLabs, Louisville, CO, 1993.
3. *EIA/IS-23 RF Interface Specification for Television Receiving Devices and Cable Television Systems*, 1997. Electronic Industries Association, Engineering Department, Arlington, VA. Available from Global Engineering Documents, Englewood, CO, 800-854-7179.
4. David Large, Defining "Cable-Ready" Consumer Equipment. *CED*, April 1993, pp. 84–90.
5. *Cable Television Consumer Protection and Competition Act of 1992*, House of Representatives Report 102-862, 102nd Congress, 2nd Session.
6. Walter Ciciora, The 1992 Cable Act and the Consumer Electronics Interface, *1994 NCTA Convention Technical Papers*. NCTA, Washington, DC.
7. The response of Intermedia Partners, authored by David Large, cochair of the IS-23 (released as EIA-23) Subcommittee, is summarized in: David Large, op. cit.
8. Summarized in: Joseph Van Loan and David Large, Tuner Characteristics of Cable-Ready Receivers, *1994 NCTA Technical Papers*, NCTA, Washington, DC.
9. CFR 47 §15.118.



Chapter 25
Home Networks

25.1 Introduction

Home networks have become commonplace since the end of the 1990s. The first home networks, which started with an Ethernet hub connecting two PCs to a cable modem, have developed to include a sophisticated residential gateway, with routing, network address translation, and firewall capabilities. In addition, a host of options has been developed to connect devices within the home, including phone line, power line, coaxial cable, and wireless alternatives.

This chapter will focus on the concept of a residential gateway designed to support high-speed data applications. A wide range of devices that support the functions of a residential gateway are already available at retail from a large number of different manufacturers, and the residential gateway is already becoming commonplace in the subscriber's home. The next trend may be to incorporate media center functions into the residential gateway, including a digital set-top terminal and a digital video recorder.

As broadband Internet access is offered to the customer, many cable operators have seized on the idea that home networking can be offered to the customer as an additional service. Home networks allow the customer to connect multiple PCs to a single broadband connection and potentially provides the cable operator with an additional revenue stream from equipment lease and a home network support contract.

This chapter will start by discussing the applications that home networks can support and the challenges they present to the cable operator. This leads us to a set of requirements for the residential gateway (RG), a network element that supports a home network and connects it to the outside world via the broadband connection. The RG binds the various aspects of the home network together. A wide range of technologies are required to support its operation; each of these is discussed in turn before we examine the CableHome initiative against the model we have developed.

Note: Much of the discussion in this chapter relies on concepts introduced in Chapter 5. We recommend you read that chapter first, especially Section 5.1, "Protocol Issues," and Section 5.3, "Non-DOCSIS Protocols."

25.2 Applications

This section describes a range of applications for home networks. This will always be an incomplete list because new applications are appearing rapidly. Although some home networks are application specific, for example, a cordless phone, others are multiservice networks, supporting data, voice, and video applications.

25.2.1 File and Printer Sharing

Personal computer (PC) home networks first became popular when a significant number of households acquired a second PC. Suddenly it was convenient to share a printer by networking it, and moving files back and forth between PCs was simplified by a home network. The technology was a straightforward adaptation of local area network (LAN) technology from the business environment. Business LANs have been commonplace for some time, and their maturity allowed most functions to be reduced to a single chip, bringing the cost reduction necessary to sell to the home market.

25.2.2 Broadband Connection Sharing

The widespread adoption of broadband Internet connection services (see Chapter 4) made PC home networks even more popular (some would say essential!). The significant monthly outlay for a broadband connection (about $35–45 per month) can be shared across multiple PCs. Many families now own two, three, or even four PCs, and a single broadband connection can be shared by parents and children for business, school, and personal use. A number of related applications that may be associated with broadband connection sharing:

- *Personal firewall* — This provides blocking of potentially harmful traffic, preventing it from reaching the devices connected to the home network.
- *Parental controls* — This provides blocking of content and access to Web sites based on profiles defined by the customer (usually the parent).
- *Network backup* — This allows customers to back up their data to remotely located storage for later retrieval in the event of a local disk problem.

Although multiple PCs are the primary driver, a broadband connection may also support connectivity for a range of other in-home devices:

- Game consoles such as Microsoft X-box, Sony PS-2, and Nintendo GameCube are all designed to operate over a broadband connection. The application is multiplayer network games, for example, X-box Live.

- Digital audio jukeboxes (HP MediaCenter) are designed to share a broadband connection to access Internet radio and CD database.
- Digital video recorders (SONICblue, TiVo) may use a broadband connection to access program guide data. Digital TVs may also offer this functionality.
- Meter reading, home security, medical monitoring, and fire alarm systems may utilize a broadband connection.

25.2.3 Video Distribution

Cable operators are very familiar with home networks. After all, they have been installing and maintaining video distribution home network for their subscribers for many years. At their most basic, video distribution networks may consist of coaxial cable and splitters. However, larger houses may require a broadband amplifier to boost the signal if cable runs are long and there are many cable outlets.

The emergence of the digital video recorder, which is capable of storing compressed digital video on a hard disk drive, is driving the need for baseband video distribution over the home network to different television displays around the home. In addition, the OpenCable effort (see Chapter 3) will make it possible to integrate the functions of a digital set-top terminal into the residential gateway.

25.2.4 Telephone

Telephone operators have typically played a role as the other traditional home network provider for analog telephone services. Traditionally, the requirements for telephony service are satisfied by a twisted pair daisy-chained around the various wall jacks in the house. The same wires are now being used for other applications, such as data transmission (see Section 25.10.2).

Wireless (or cordless) phones are commonplace and may operate in several frequency bands, the most common being 800 MHz, 2.4 GHz, and 5.8 GHz. Wireless phones operating in the same band and the same location as a wireless home network may cause interference, resulting in reduced throughput and significant packet loss.

25.2.5 Home Security and Management

Traditionally, home security systems have used a telephone modem to dial the service provider in the event of an intrusion or fire alarm. This is no longer necessary with access to a broadband connection. Moreover, always-on connectivity will allow the service provider to extend service to include video surveillance. In addition, demand-side power management can be supported. These systems enable the power company to implement demand-side power shedding, allowing them to reduce the cost of peak power generation.

25.3 Challenges

Home networks bring additional challenges in the areas of installation, support and maintenance, security, and billing.

25.3.1 Installation

Installation and support of a home network brings many issues for the cable operator.[1] Most homes are not cabled with Category 5 wiring suitable for Ethernet connections (although some progressive builders are installing these cables as an option in upscale developments[2]). Installation of new wiring is rarely cost effective, so most installations focus on the use of existing wiring or on wireless network technologies.

New equipment must be configured correctly for it to operate, and this is often a nontrivial exercise (as anyone who has installed a home network will attest). Most home network equipment requires software to be installed on at least one personal computer to allow configuration of the residential gateway.

25.3.2 Support and Maintenance

A home network introduces new wiring, passive devices, and active devices into the service equation. The wiring itself may be a cause of ongoing support issues unless it is installed correctly, leading some operators to offer support only for networks that they have installed (or at least inspected). In most home networks, there is at least one active device, typically powered by an AC adapter. The customer may need to be educated to check the power to that device before placing a service call. Fortunately, most devices are equipped with nonvolatile memory so that configuration information is not lost due to power interruptions. Nevertheless, the additional components in the home network add additional failure modes to the service equation. Many cable operators may not even be aware that a customer is operating a home network and do not track this information in their support systems. It may be difficult to determine whether the home network or the cable network is responsible for the outage, making fault finding difficult. Ideally, customer service agents should have real-time access to diagnostic tools, allowing them to localize a fault during a service call.

Another huge challenge to the cable operator is to gain "visibility" to those devices that reside behind the residential gateway for troubleshooting purposes. For example, a game console that does not contain the typical suite of network troubleshooting tools (such as "ping" and "traceroute") certainly complicates matters. (See Section 25.12.1)

25.3.3 Security

Home networks have a significant effect on the security of the service, especially for IP-based services such as high-speed data and voice over IP. There are

several aspects that are of concern, including protection from Internet-originated attacks, theft of service, and privacy.

Internet-originated attacks can take a wide variety of forms, but a broadband connection exacerbates the problem by allowing faster communication speeds as well as by being always connected. For this reason, many residential gateways include additional security protection measures.

The privacy of the broadband connection is protected by encryption between the CMTS and the DOCSIS modem (see Chapter 5, DOCSIS Security). However, packets exchanged between the cable modem and the PC are not encrypted. This is not a problem if a physical wire is used. But in the case of a wireless network, the customer's data is being broadcast in the free air. It is relatively easy to intercept that customer's data with a wireless receiver to find passwords, bank accounts numbers, and other personal data.

A customer may unknowingly share a broadband connection by using a wireless local area network (LAN) without enabling encryption. (Most products are shipped as open systems by default.) Wireless networks can operate over a range of several hundred feet, so a neighbor may stumble across an active wireless network and share the customer's broadband connection. In other cases, deliberate theft of service may be attempted via wireless technology.

25.3.4 Billing

If the cable operator decides to sell a home networking service, then a simple monthly billing option can be implemented quite easily. However, if the customer does not pay the bill, a method is required to remotely disable the service without rolling a truck.

25.4 Requirements for a Residential Gateway

In the discussion so far, we have demonstrated the breadth of the applications and technologies that come under the umbrella term of *home network*. The rest of this chapter will focus on *home data networks* because they are the technology most likely to be encountered by the cable installer in the immediate future. Most cable operators are initially focusing on home data networks because the technology is reasonably mature and because they offer the customer the best value proposition. In addition, more products are available at retail, and many subscribers have already experimented with the installation of their own home data network. This also means subscribers will be more aware of home data networks than of other types. All of this adds up to a near-term opportunity for cable operators to generate additional revenue by providing installation and support services for home networks.

At this point, it is helpful to define the concept of a *residential gateway* (RG). As defined in Chapter 5, a gateway is a router that interconnects two or more

networks — in this case the wide area network (WAN) to the local area network within the home. As shown in Figure 25.1, the RG terminates the broadband connection into the home and supports a home network, allowing various devices to be interconnected (in this example, including personal computers, a telephone, a game console, and home security system). The RG may also include the DOCSIS cable modem functions and an embedded media terminal adapter (E-MTA) for voice over IP support. In the near future, the RG may also support video functions.

Now that we have defined the residential gateway, we can discuss the requirements that it must meet.

Connectivity — The RG must connect the cable modem to various PCs, printer(s), game consoles, and any other devices, preferably without the need to install any new cables in the home.

Multiplexing and address management — The RG must support the flow of packets between the various devices within the home while also allowing them to share a broadband connection. In order to support packet forwarding at the network layer, the RG must assign all devices a unique IP address.

Configuration and performance management — The RG must operate with minimal configuration. It should support local and remote configuration. In addition, the RG must provide some indication of the performance and

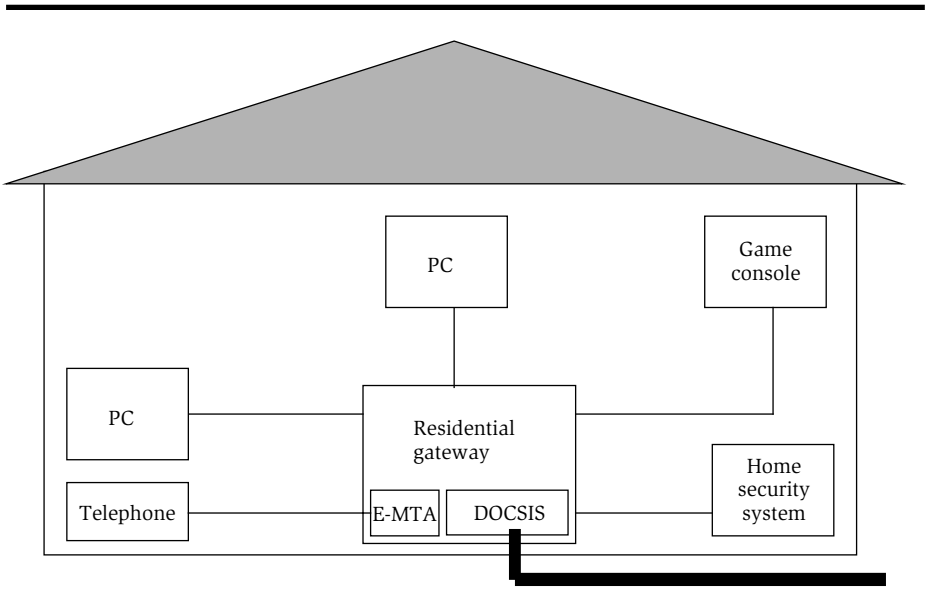


Figure 25.1 Residential gateway.

operational metrics of the home network to support troubleshooting. The RG must be remotely software upgradable by the service provider.

Security — The RG must provide mechanisms to prevent unauthorized access to the home network. It must also prevent packets that are exchanged between various devices within the home from being forwarded to the wide area network.

Quality-of-service (QoS) — The RG must minimally provide a best-effort packet delivery service between the various devices it connects, which is adequate for high-speed data applications (for example, Web browsing and email). However, QoS guarantees will be required to support reliable, high-quality voice and video services. QoS is specified by setting limits for maximum packet loss, maximum packet delay, and maximum packet delay variation (jitter).

How the RG satisfies each of these requirements will be discussed in more detail in the next sections.

25.5 Connectivity

A residential gateway must be able to connect the various devices together at a data rate that is sufficient to support the intended applications.

Data rates of up to 100 Mbps are possible over Category 5 cables using 100-base-T (Fast) Ethernet. However, most PCs are not able to support data transfers at these rates, and if the PC is accessing the Internet through a broadband connection the rates will typically be rate shaped by the cable operator to 1.5–2 Mbps downstream (from the network) and 256–384 kbps upstream. Therefore, although higher rates are preferred, 2 — 5 Mbps is usually sufficient for data applications within the home network.

The RG must be able to support a range of physical media, depending on the type of media available in the home. The media may be physical wires (Category 5 cables, home phone wiring, power line, or coaxial cables — see Section 25.10), or wireless networking may be preferred for mobility or ease of installation (see Section 25.11)

25.6 Multiplexing and Address Management

In the early days, most home networks were constructed by connecting several PCs to an Ethernet hub. If a broadband connection was available, the cable modem was also connected to the hub, allowing several PCs to access the broadband connection simultaneously. However, since the hub is a transparent (layer 2) device, the operator would have to allocate an additional IP address to each PC added by the customer. The DOCSIS cable modem includes a configuration

parameter to limit the number of IP addresses than can be allocated; by default, most cable operators set this value to 1. Therefore, when a customer connected a second PC, it would not receive an IP address. After a call to the help desk, the operator would increase the maximum IP addresses parameter for that cable modem and everything would start to work. Many operators generate a significant revenue stream by charging $5–10 per month for additional IP addresses. However, the hub provides no filtering of packets, so Ethernet frames addressed from one PC to another are also forwarded by the cable modem to the wide area network. This is a problem because (a) it causes additional traffic over the broadband connection and (b) many customers left their PCs in their default, unprotected configuration, sometimes unknowingly sharing the entire contents of their PC with anyone else on the network! For these reasons alone, a residential gateway may be considered necessary by both the customer and the cable operator.

The residential gateway must multiplex the various packet flows so that they can coexist over the various shared media within the home network and so that all devices can share a single broadband connection simultaneously. Since public IP addresses are becoming scarce, the RG must support multiple devices behind a single IP address.

Multiplexing can take place at the data link, network, or transport layer. In practice, a combination of these techniques is often employed by the RG, as illustrated in Figure 25.2. The Ethernet ports and the 802.11b wireless ports (see

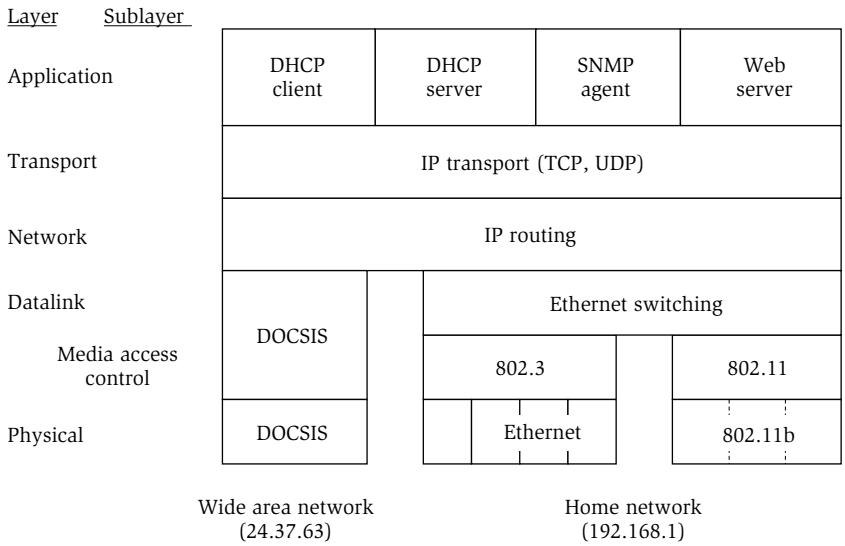


Figure 25.2 Residential gateway protocol stack.

Section 25.11) are interconnected at the data link layer using an Ethernet switch. This allows PCs and other devices to communicate at the data link layer within the home network. In addition, a routing function is used to connect two different networks, labeled wide area network and home network. Moreover, in order to support multiple devices behind a single IP address, transport layer multiplexing is also required (see Section 25.6.2).

The various devices operating at the TCP-IP level must be assigned a unique IP address on the network before they can exchange packets. As shown in Figure 25.2, the RG contains both DHCP client and DHCP server functions to support address assignment. (The DHCP protocol is described in Chapter 5.)

The DHCP client in the residential gateway obtains an IP address from the cable operator's DHCP server and assigns it to the DOCSIS physical interface. From a network point of view, the RG behaves exactly as if it were a directly attached PC. In our example, the network number is 24.37.63 and the RG is assigned a DHCP lease for the public routable address 24.37.63.12. (Most cable operators assign their subscribers a public IP address, although at least one operator is assigning private addresses and performing address translation of those private addresses to a pool of public addresses on the wide area network side of the CMTS.)

The residential gateway needs a second IP address in order to communicate on the home network. However, only one public address has been assigned by the cable operator to this subscriber, and that has already been allocated. Therefore, the RG assigns a private, nonroutable address to its physical interface on the home network. In our example, the RG uses a private Class C network of 192.168.1 and assigns itself an IP address of 192.168.1.1. (Note that 192.168.1.1 through 192.168.1.254 is an arbitrary nonroutable address range for the home network, specific address ranges vary by manufacturer and are usually configurable.)

As PCs and other devices are connected, they will make DHCP requests to the residential gateway. These are fielded by the RG's DHCP server, and each device on the home network is assigned a unique IP address from the private Class C network. In our example, assignments start at 192.168.1.100 and increase by 1 each time. The DHCP server also informs each client that it is the default gateway (192.168.1.1) for the home network, so all messages that need to be routed are directed to it. In addition, the DHCP server passes along the domain name server (DNS) addresses, which it obtained from the cable operator's DHCP server. The final network topology, including address assignments, is shown in Figure 25.3.

25.6.1 Network Address Translation (NAT)

As the residential gateway forwards packets from the home network to the wide area network, it translates the source address of the packet from the private address of the source device to the public address of the RG. This process is called *network address translation* (NAT).[3]

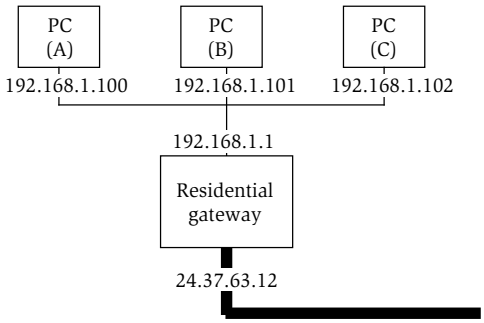


Figure 25.3 Network address translation.

In Figure 25.3, the RG has assigned each PC a unique IP address, starting at 192.168.1.100, which is a private Class C address. As you can see from this example, these private addresses typically bear no relation to the public address assigned to the WAN port (24.37.63.12). NAT effectively decouples the WAN IP addressing from the home network IP addressing. As packets are routed from the PC to the WAN, the RG replaces the private source IP address with its public source IP address (as shown in Figure 25.4). This is necessary because the

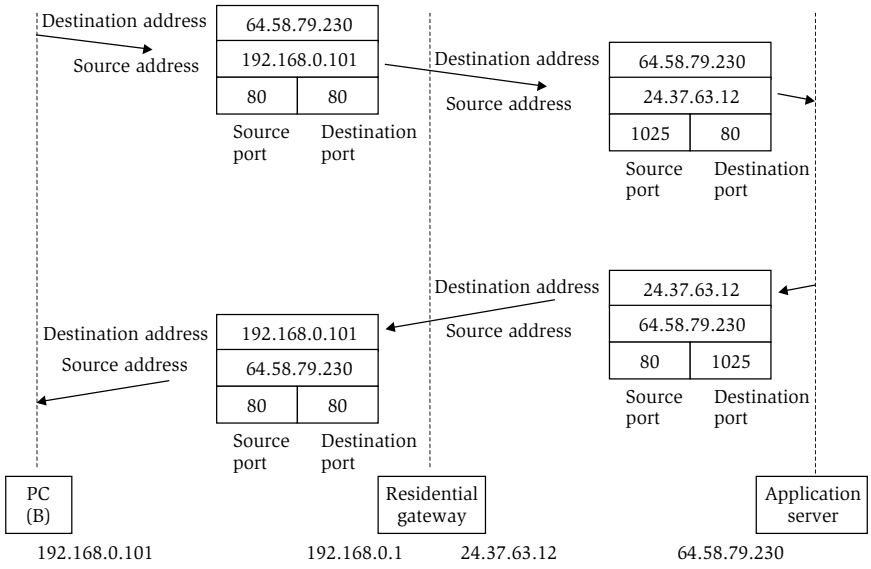


Figure 25.4 Network address port translation.

private IP addresses are unroutable over the Internet. In this way, the far-end application will send response packets back to the RG. When the response to the packet is returned from the WAN, the public destination IP address of the RG is translated back to the private destination address of the PC. However, the RG has to forward incoming packets from the Internet to only one of the PCs on the home network. How does it choose which one? The solution is provided by network address port translation.

25.6.2 Network Address Port Translation (NAPT)

Network address port translation (NAPT), a special type of NAT, is commonly used in the residential gateway to allow a single public address to be shared by several hosts (PCs or other network devices). Figure 25.3 illustrates a typical arrangement where an RG performs NAPT. The RG has to choose to route an incoming packet to only one of the three PCs, the one from which the original request came. To do this, the RG must maintain the state of each IP flow and multiplex packets at the transport layer.

As explained in Chapter 5, the Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) maintain endpoint identifiers at the transport layer called *ports*. The RG creates a table of port translations (shown in Figure 25.5) so that each IP flow is identified by assigning it a unique source port number. As the packet is routed from the PC to the WAN, the original source port number is replaced. Figure 25.4 illustrates an example where PC (B) sends an HTTP request to an application server. The original packet has a source port number of 80, which is translated to a source port number of 1025 by the RG. As the packet is routed, the RG dynamically creates an entry in the address and port translation table in the RG (see Figure 25.5, second row).

Source PC	IP Address	Source Port Number	WAN Address	Source Port Number
A	192.168.1.100	25 (SMTP)	24.37.63.12	1024
B	192.168.1.101	80 (HTTP)	24.37.63.12	1025
C	192.168.1.102	80 (HTTP)	24.37.63.12	1026
C	192.168.1.102	21 (FTP control)	24.37.63.12	1027

Figure 25.5 Address and port translation table.

When the packet reaches the application server at the far end of the connection, it takes the source port number and places it in the destination port number field in the response packet. When the RG receives the response packet, the destination port number (1025) identifies the PC that made the request. The original port number (80) is restored before it is relayed to the PC so that it remains unaware of the port number substitution. Thus, multiple PCs and other network devices can use the same application behind a single WAN address without confusion.

Most of the time, network address port translation does not affect the applications at the endpoints of the network. Nevertheless, it is important to understand that NAPT violates an important Internet principle, called *end-to-end network transparency*. For example, File Transfer Protocol (FTP) uses multiple ports for a single transaction — one for control and others for data transfer. To support FTP, an application-level gateway (ALG) must be built that associates the multiple port numbers used by FTP and translates FTP-specific fields in the control messages so that they refer to the translated port numbers. The NAPT function in the residential gateway generally incorporates application-level gateways for all common protocols that need this type of special attention. The longer-term problem is that new Internet protocols are being developed all the time, and new application-level gateways may have to be built to support some of them.

25.7 Configuration and Performance Management

An important goal for a residential gateway is that it works with minimal user configuration — the ideal is a *plug-and-play* device that autoconfigures the necessary parameters. Nevertheless, some configuration is often required to allow for different broadband connection types and to provide adequate security within the home network environment.

Business LANs are typically supported by an information technology (IT) department, but home networks must be simple and easy to set up and maintain without onsite, professionally trained staff. Many MSOs are starting to provide a managed home networking service to their customers as part of the high-speed data bundle. To support these customers, remote configuration management is an important tool to reduce operational costs.

At a minimum, the residential gateway should be configurable by using Simple Network Management Protocol (SNMP) commands from a network management system (NMS). Figure 25.2 shows an SNMP agent that can accept configuration management requests from the wide area network (remotely) or the home network (locally). In addition, a Web-based user interface is invaluable for initial configuration, and the RG typically incorporates a Web server that allows direct configuration from any Web browser running on a local PC.

The residential gateway should support Trivial File Transfer Protocol (TFTP[4]) to load new versions of firmware.

Performance management functions must be supported to allow troubleshooting of the home network. The SNMP agent and Web server in the RG should support status reporting to allow the cause of a failure to be identified. However, diagnosing problems in a badly configured TCP-IP network can be challenging even for a knowledgeable service technician; therefore, the SNMP agent should support remote performance management capabilities. Remote performance management is invaluable because it allows a customer service representative to get real-time information about the performance of the customer's home network during a trouble call.

25.7.1 Simple Network Management Protocol (SNMP)

In home networks, Simple Network Management Protocol (SNMP) is commonly used to ease management and configuration of the network devices (the original version of SNMP was developed to manage routers). The current, third version of SNMP is called SNMPv3.[5] Although there is space here for only a brief overview, many resources exist for those readers who would like to learn more about SNMP.[6–8]

SNMP is designed to allow a management system to access defined parameters that affect or reflect the state of a device. The management system communicates directly with an SNMP agent that is implemented in software or firmware on the managed device. SNMP allows a parameter to be written using a "Set" command and allows a parameter to be read using a "Get" command. Commands are sent and responses received on User Datagram Protocol (UDP) port 161. In addition, SNMP supports an event notification mechanism, allowing the SNMP agent to send an unsolicited "trap" message on a predefined event (for example, an interface that changes state). Traps are sent on UDP port 162. Although the protocol itself is simple, a general data-modeling language called the structure of management information (SMI) allows a sophisticated data model to be constructed for each type of managed device. SMI uses abstract syntax notation 1 (ASN.1) to define a management information base (MIB) that is completely independent of the SNMP message structure.

Although a unique "enterprise" MIB could be defined for each class of device, it is generally not necessary. Over 100 standard MIB definitions already exist. The MIB is a hierarchical, treelike definition that allows a device to be logically broken down by function so that common functions can share the same definitions. Every SNMP-manageable device must support the MIB-2 definition, illustrated in Figure 25.6. For example, the system description of the device can be read by issuing an SNMP "Get" command and specifying the field as "iso.org.dod.internet.mgmt.mib2.system.sysdesc" (for compactness, the index of each field is transmitted in SNMP, "1.3.6.1.2.1.1" in this example.) The text string returned by the residential gateway might read "Residential Gateway, Hardware Version 1.0, Firmware Version 1.2.1, 'Manufacturer Name,' 'Model Number.'"

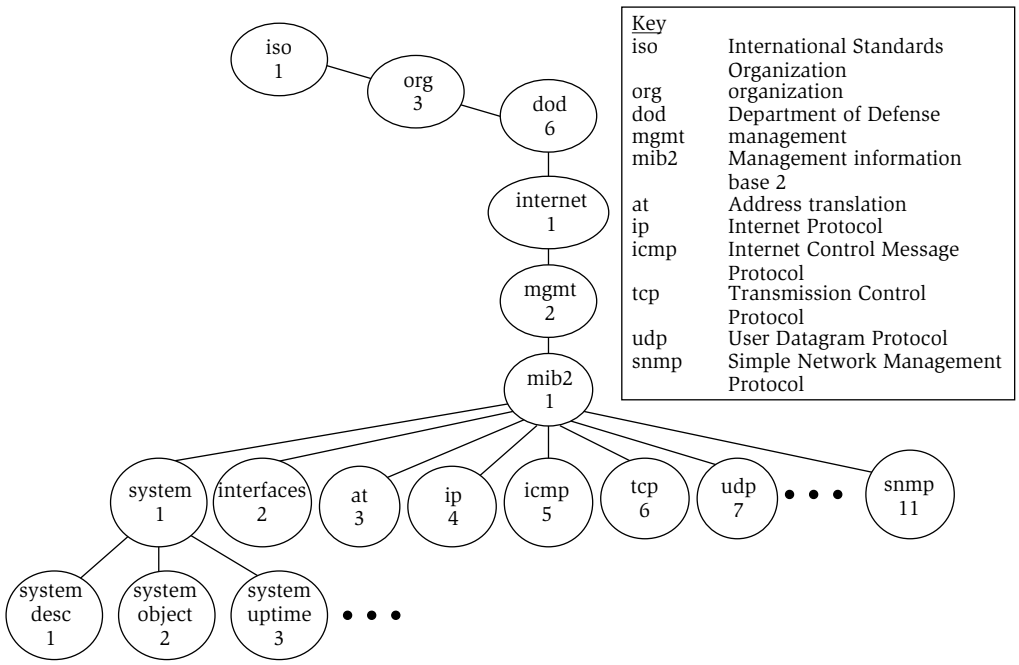


Figure 25.6 MIB-2 structure.

Until fairly recently, SNMP has been used primarily for performance monitoring because of security issues in the earlier versions of the protocol. SNMPv3 adds authentication and encryption to the protocol, making it considerably more useful for configuration management.

25.8 Security

The RG must provide mechanisms to prevent unauthorized access to the home network. Since most attacks originate from the wide area network (Internet) side of the RG, the main requirement is to prevent suspicious IP packets from being forwarded to the home network. Most businesses protect their corporate LANs from unauthorized access with some type of firewall.[9] A *firewall* is a robust IP router that prevents certain IP packet flows that are known to be security hazards. The home network needs to be protected from Internet-based attacks, especially in the case of an always-on, broadband connection such as a cable modem. The RG is the logical point to incorporate a firewall.

25.8.1 Firewall

A firewall is a device that provides protection against unauthorized access to a network. Just as a firewall protects the occupants of a car from an engine fire, a network firewall protects the network from the outside world. Figure 25.7 illustrates the operation of a firewall. The firewall allows authorized flows, numbered 1 and 2 in the diagram, but blocks unauthorized flows, numbered 3 and 4. The rules for blocking can be provisioned by configuration management messages from a management device (in this case the PC that is being protected by the firewall).

Firewalls may be implemented in software on a PC or may be a separate device. Firewall functions are often integrated into a router or gateway because a separate device can be hardened to be more robust in the presence of network attacks. For example, a common attack is called a *denial-of-service* (DoS) attack, where the attacker focuses a large volume of traffic on the device under siege. The intent of a DoS attack is to overload the host so that a weakness is exposed. A separate hardware device can be specifically engineered to withstand these types of attack. Moreover, if the firewall is integrated into a router or gateway, it can be used to protect a number of hosts.

A number of different techniques are implemented by firewalls. Some of the most important of these are network address port translation, inbound request blocking, and stateful packet inspection.

- *Network address port translation* (NAPT) — Strictly speaking, the primary purpose of NAPT is to expand the address range within the home network without consuming the scarce resource of public IP addresses (see Section 25.6.2). Nevertheless, NAPT provides considerable protection from malicious network traffic by hiding the private addresses behind the firewall. Any traffic from the WAN is directed at the public address rather than

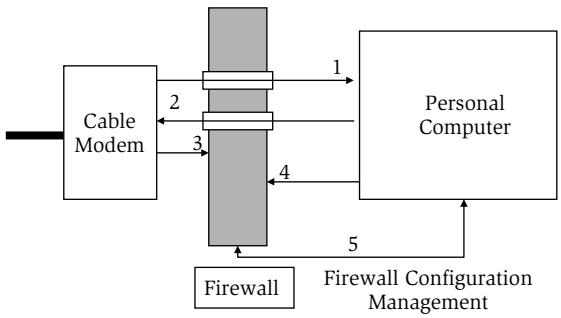


Figure 25.7 Firewall operation.

private addresses. Hardware devices can be much more robust against network attacks by avoiding security weaknesses that exist in the PC operating system that are exploited by such attacks.

- *Inbound request blocking* — A common attack is called *pinging*. In this, the attacker tries to determine unprotected hosts (typically, the subscriber's PC) that exist on the network. The attacker will cycle through a series of addresses, sending a ping message for each, until a response is received. At that point, the attacker has the IP address of a potentially vulnerable host and can then proceed to attempt to break into that host. For a typical subscriber, most inbound packets can be discarded without any side effects, preventing the attacker from being able to determine that the host exists and cutting short a potential attack.
- *Stateful packet inspection* — This is a refinement of inbound request blocking. The state of all IP flows is captured and monitored by the firewall. For example, TCP flows always start with a SYN message, and responses are acknowledged by ACK messages. By maintaining the state of all valid flows initiated by the host, the firewall is able to reject any that are not in response to a request from the subscriber. For example, a TCP SYN inbound from the network would be discarded (a common DoS attack is based on SYN flooding). When a PC initiates a TCP connection, a cache entry is created by the firewall. Only incoming TCP traffic that is consistent with a TCP connection established from within the LAN will be forwarded. Similar techniques can be used to restrict UDP and ICMP traffic.[10]

25.9 Quality of Service

Today, quality of service (QoS) is defined as best effort for broadband data services. However, voice and video services are much more sensitive to loss, delay, and jitter.[11] Considerable effort is being spent on developing higher-bit-rate solutions with provision for reliable packet delivery so that compressed digital video can be distributed within the home. Ensuring QoS over data networks is a subject of considerable complexity.[12]

25.10 Wired Networks

Business LANs typically use 10-base or 100-base Ethernet connections (see Chapter 5), which require a 2-pair, Category 5 cable. However, such cabling is rarely present within the typical home environment and new cables are costly to install. Therefore, home network equipment manufacturers have adapted the Ethernet physical layer to support various physical media, including phone lines, power lines, and coaxial cables. All of these technologies are proprietary, and

detailed technical specifications are made available to manufacturers only under nondisclosure agreements. Therefore, the reader is referred to the manufacturers' literature for detailed installation and operating instructions of wired home networking solutions.

25.10.1 Ethernet

If the home is already wired with the necessary Category 5 cable, traditional 10-base-T or 100-base-T Ethernet is an excellent home network solution. As a rule of thumb, any devices located in the same room should use Ethernet as the preferred solution. For example, a home office may be the location for the residential gateway, a PC, and a print server, allowing them to be connected directly with standard Ethernet cables (see Chapter 5). For this reason, most residential gateways typically have up to four Ethernet ports as standard.

25.10.2 HomePNA

HomePNA is an alliance of over 150 companies (www.homepna.org) that support a single technology for communication over existing unshielded twisted-pair phone wiring within the home. The current version of the specification, HomePNA 2.0, is based on technology from Broadcom and Lucent Microelectronics. HomePNA 2.0 operates in a 4- to 10-MHz band with a symbol rate of 2 or 4 Msps. The range is quoted as 1,000 feet.

One of the major challenges of using existing phone wiring is that it can impose significant attenuation at certain frequencies within the chosen band (so called *nulls*). To accommodate for channel impairments, the modulation is adaptively varied from QAM-4 through QAM-256 achieving from 2 to 8 bits per symbol. In addition, a technique called *frequency diverse QAM* (FDQAM) is used to send two duplicate copies of the data signal so that they are spread out across the frequency band such that a null in the band does not completely block the signal. If the passband has reasonably flat spectral characteristics, the system will adapt to operate at twice the symbol rate (4 Msps) using standard QAM modulation, allowing a data rate of up to 32 Mbps.[13]

25.10.3 Powerline

The HomePlug Powerline Alliance (www.homeplug.org) is an alliance of approximately 70 companies. The latest version, HomePlug 1.0.1 (often abbreviated to 1.0), was developed by Intellon.

The major challenge in using power lines is the high levels of noise that can be injected into them by home appliances (for example, hair dryers, electric drills). To combat noise, HomePlug uses orthogonal frequency division multiplexing (OFDM) across 84 carriers between 4.3 and 20.9 MHz. The system adaptively selects which carriers to use based on interference. The modulation

scheme is differential quadrature phase shift keying (DQPSK), and several forward error correction techniques are used. When all 84 carriers are simultaneously active, a raw data rate of 20 Mbps can be achieved. After overhead, the maximum user data rate is 14 Mbps.

56-bit DES-based encryption is standard, mainly to prevent accidental interconnection between customers sharing the same power transformer. HomePlug includes a key distribution protocol. Unfortunately, encryption does not address the problem of contention, which might significantly reduce throughput if several home networks are operating on a common power line.

25.10.4 Coaxial Cable

The coaxial cable within the home used for distribution of broadband television signals would seem to be an ideal medium for data networking. For the short distances within the home, the frequencies above the highest cable channel (typically 750 MHz and in some cases 860 MHz) are quite usable. Typically, frequencies above 900 MHz have been used by some proprietary systems. However, the coaxial cable in older houses is not typically routed to all rooms, and a coaxial outlet may not be sited conveniently.

A recent suggestion has been that the 2.4-GHz spectrum of 802.11b and 802.11g may be carried quite effectively by in-home coaxial wiring. At these frequencies, the signal is severely attenuated by the cable itself and any splitters in the path, but the propagation through the coaxial cable is still superior to free air. Moreover, the cable provides shielding from interference.

25.11 Wireless Networks

Wireless networks are very similar to conventional wired local area networks, in the sense that they work by exchanging Ethernet frames between endpoints. There is not space here to provide more than an overview of the most important concepts of wireless networks, but there are excellent texts available for further reading.[14,15] Although a number of standards exist for home networking, we will focus on the 802.11 family (also known as WiFi) because they are best suited for high-speed data home networks.[16]

The 802.11[17] standard has three common variants: 802.11b and 802.11g operate in the 2.4-GHz industrial, scientific, and medical (ISM) band, and 802.11a operates in the 5-GHz unlicensed national information infrastructure (U-NII) band. A typical 802.11 wireless network is shown in Figure 25.8. The endpoints are called *stations*; the hub of the network is a special class of station called an *access point*. The access point provides connectivity to the wired network. When a station is within range of an access point, it may *associate* with it. Once associated with an access point, stations may exchange packets with it but not directly with each other.

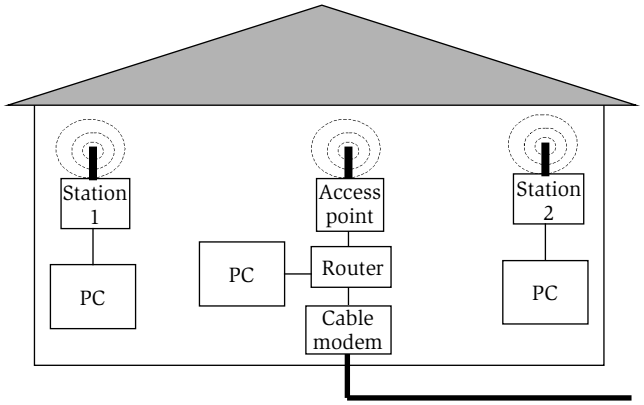


Figure 25.8 Wireless local area network.

Introducing a wireless medium for packet exchange introduces some additional complications for operation and management.

- A wireless network does not have a fixed boundary. In fact, the "local area" of the network may change over time as wireless propagation characteristics change. For example, even small adjustments of the antenna position of the access point can significantly affect the signal strength at the other stations. Moreover, the in-home environment can be challenging for wireless propagation in the 2.4-GHz or 5-GHz band. Walls, ceilings, and floors attenuate the signal, and any metal objects cause reflections. For these reasons, propagation characteristics are typically asymmetric — it may be possible that a station can receive the signal from an access point, but the access point cannot receive the station's signal.
- Each station in a wireless network must follow power-level management procedures to balance transmit power levels with performance. This is particularly important for battery-operated devices, such as laptop computers, in order to minimize current draw.
- The wireless bands in use are subject to interfering carriers and noise. In fact, microwave ovens may generate considerable interference in the 2.4-GHz band. Therefore, the packet transfer is inherently less reliable than over a wired network. (This is probably the reason that manufacturers recommend that you do not attempt a firmware upgrade of the station via a wireless connection!)
- As stations come within range of the access point, they may associate with it. This is equivalent to connecting a PC to a LAN hub. Stations may be mobile, and, as they move, they may associate and disassociate with an access point. Therefore, the topology of the local area network is dynamic.

- Wireless networks can be operated as *open networks*, where any station within range can join the network. This is useful for public access points, but probably not what most homeowners intend when they purchased broadband access for their family's use (and not their neighbor's). Therefore, security procedures are required; these, in turn, require configuration and management.
- Due to propagation and positioning of stations, not all of the stations may be within range of each other, and yet they may all associate with the same access point. This is illustrated in Figure 25.8, where station 1 can transmit to and receive from the access point but is unable to detect when station 2 is transmitting. For this reason, the carrier-sense function used in Ethernet networks is not always sufficient to prevent collisions; station 1 and station 2 could transmit at the same time without sensing each other's carriers, and their transmissions would collide at the access point. To remedy this situation, the 802.11 standard introduces the concept of *virtual carrier sense* that essentially makes the access point the traffic cop, distributing transmission reservation information to the stations. (The virtual carrier-sense mechanism uses a Ready-To-Send, Clear-To-Send Protocol to communicate to the stations; as such, it is less efficient than the conventional *physical* carrier-sense mechanism. Virtual carrier sense is used only when the incidence of collisions warrants it.)

Although these issues may seem overwhelming, manufacturers and standards committees have made remarkable progress in solving them. As a result, wireless products now exist to build a reasonably secure, wireless home network with relatively few operational problems. The predominant standard in the marketplace today is 802.11b. 802.11a is gaining acceptance at the time of writing, while 802.11g is newly standardized. These standards are collectively known as Wi-Fi. (The Wi-Fi Alliance is a nonprofit international association formed in 1999 to certify interoperability of wireless local area network products based on the IEEE 802.11 specification.[18])

25.11.1 IEEE 802.11

The IEEE 802.11 standard defines the media access control (MAC) and physical (PHY) layers for the multiple, overlapping, wireless LANs operating in the 2.4-GHz band. 802.11 defines two PHY layer alternatives based on either frequency hopping spread spectrum (FHSS) or direct sequence spread spectrum (DSSS).

Frequency Hopping Spread Spectrum (FHSS)

Frequency hopping spread spectrum (FHSS) is used across the 79-channel set of the 2.4-GHz ISM band at a hopping rate of 2.5 hops per second. Each channel is

1 MHz wide, and frequencies are centered at 1-MHz intervals between 2.402 and 2.480 GHz. The data rate may be 1 or 2 Mbps, corresponding to 2- or 4-level Gaussian frequency shift keying (GFSK) modulation. To allow a station to synchronize with an access point, 802.11 defines beacon frames that include information about the hop sequence number and hop index.

Direct Sequence Spread Spectrum (DSSS)

Direct sequence spread spectrum (DSSS) transmission employs a chipping code to "spread" the transmission over a wider frequency band than it would normally occupy. DSSS is a form of CDMA (see Chapter 4). But instead of a pseudorandom binary sequence (PRBS), an 11-bit Barker code is used as the spreading sequence. However, the same principles apply: The redundant bit pattern in the spreading sequence increases the signals resistance to interference and also distributes the power over a 22-MHz-wide channel that looks essentially like low-power, wideband noise. An important characteristic of DSSS modulation is that it is rejected by conventional, narrowband receivers that might be sharing the same ISM band.

For DSSS operation, the 2.4-GHz ISM band is divided into three nonoverlapping or six overlapping frequency bands, as shown in Figure 25.9. Since the

Overlapping Channel Set		Nonoverlapping Channel Set	
Channel	Center Frequency	Channel	
1	2.412 GHz	1	2.412 GHz
3	2.422 GHz		
5	2.432 GHz		
		6	2.437 GHz
7	2.442 GHz		
9	2.452 GHz		
11	2.462 GHz	11	2.462 GHz

Figure 25.9 802.11b channel definitions for North America.

DSSS channel occupies a 22-MHz band, only the nonoverlapping channels are guaranteed to be free from contention with each other. For example, two wireless stations operating on channels 1 and 3 would not associate, but carrier sensing may prevent them from transmitting packets simultaneously.

Wired Equivalent Privacy (WEP)

In addition, 802.11 specifies optional wired equivalent privacy (WEP) to provide security from eavesdropping and unauthorized access. WEP uses a shared-secret key, which is 40 bits in length. The shared-secret key is concatenated with a 24-bit initialization vector (IV) to create a 64-bit key, which is used to encrypt packets according to the RCA RC4 PRNG algorithm. The IV is a pseudo-random number, sent in the clear in the packet header. Many manufacturers have extended the shared secret to 104 bits in length, which is combined with the 24-bit IV to create a 128-bit key. However, there is no guarantee of interoperability between different manufacturers when operating with 128-bit encryption.

Unfortunately, 802.11 does not specify a key management protocol, so setting up the shared secret involves typing the same hexadecimal string into the access point and station setup screens. For 40-bit keys, a 10-digit hexadecimal string must be entered at every station on the network. For 104-bit keys, a 26-digit hexadecimal string is required.

Moreover, there have been successful attacks on the WEP encryption, and certain shared keys have pronounced weaknesses. A software tool is readily available that can crack a typical WEP-protected LAN in 2–3 hours. Nevertheless, for a typical home user, WEP still provides a basic level of security.

WiFi Protected Access (WPA)

The shortcomings of WEP have been recognized for some time. A new encryption mechanism, called *WiFi protected access* (WPA), developed by the WiFi Alliance, became available in April 2003. WPA is a subset of the 802.11i security standard.

25.11.2 IEEE 802.11b

IEEE 802.11b[19] has become quite popular, with approximately 15 million units sold as of the end of 2002. IEEE 802.11b extends the 802.11 MAC to higher data rates by the addition of a new physical layer called high-rate DSSS (HR-DSSS). The maximum data rate is increased to 11 Mbps by using *complementary code keying* (CCK) modulation while maintaining the same chipping rate of 11 MHz. Two higher rates are added, 5.5 Mbps (4 bits per symbol) and 11 Mbps (8 bits per symbol), using a set of 64 eight-bit complementary code words as the spreading sequence. The code words are chosen to make them easily discriminated by the receiver, despite noise and multipath interference.

Operational Guidelines

Configuration of an 802.11b wireless network is relatively straightforward. The access point (or wireless base station) is usually installed first. No configuration is strictly necessary because factory defaults will usually work. However, several important parameters should be checked and configured.

1. *Channel* — 802.11b access points are usually set to transmit on channel 6 by default, and this channel can become quite congested if there are other wireless networks operating nearby. A simple check is to scan for networks using the tools that are supplied with the access point and to choose a channel that is not in use. For example, channel 1 or channel 11.
2. *Service set identifier* (SSID) — The SSID is a unique character string (up to 32 characters in length) that identifies the wireless network. Most access points will have a default string, for example, "default." Changing this to a meaningful string is an important step that will help with troubleshooting should it be necessary later.
3. *Wired equivalent privacy* (WEP) — By default, access points are shipped as "open" networks; that is, encryption is not enabled. At a minimum, a 40-bit WEP shared key should be selected and WEP encryption set as "mandatory."

Once the access point is installed and configured, the other stations can be set up. It is important to make sure they are configured to associate only with the SSID selected in step 2. By default, a wireless station will scan for any home networks that it can detect and prompt the user to select one. Unless an SSID is specified, it is easy to accidentally associate with the wrong home network, and this can cause endless confusion. Finally, the shared WEP key should be entered and WEP encryption set to mandatory for each wireless station. As a final installation check, the signal strength at the access point and each station should be checked with the tool provided by the manufacturer. It should be good or excellent — anything less than this should be remedied by moving the stations (sometimes mounting the station on a shelf rather than a metal desk can make a considerable difference in signal strength). The customer should be educated not to move the wireless stations unnecessarily and cautioned to check the signal strength if they do. If the signal strength is not good to excellent, the performance of the network may be compromised — just like a telephone modem, the wireless stations will "fall back" to a lower-order modulation, reducing data throughput.

These operational guidelines apply equally to 802.11g, which will be described next.

25.11.3 IEEE 802.11g

IEEE 802.11g extends the data rate of 802.11b from 11 Mbps to 54 Mbps while remaining backward compatible with it at lower data rates. The standard was

approved in June, 2003. Using the same 2.4-GHz ISM band, 802.11g employs the same *orthogonal frequency division multiplexing* (OFDM) modulation scheme as 802.11a to achieve the higher data rates. (Until recently, the FCC prohibited the use of OFDM in the 2.4-GHz ISM band.[20]) In addition, 802.11g includes two other optional modulation schemes: CCK-OFDM and packet binary convolutional coding (PBCC).

25.11.4 IEEE 802.11a

The IEEE 802.11a[21] physical layer is designed for operation in the 5-GHz unlicensed national information infrastructure (U-NII) bands. The channels assignments and allowed power levels are shown in Figure 25.10. The 5-GHz band is less prone to interference than the 2.4-GHz band. The U-NII lower band is intended for short-range applications, the middle band for midrange applications, and the upper band for longer-range, point-to-point applications using directional antennas at the endpoints.

Orthogonal frequency division multiplexing (OFDM) is used to spread the payload across 48 data subcarriers (see Chapter 4 for a description of OFDM principles).

Frequency Band	Operating Channel Numbers	Channel Center Frequencies (MHz)	Maximum Output Power (with up to 6 dBi antenna gain)
U-NII lower band (5.15–5.25 GHz)	36 40 44 48	5,180 5,200 5,220 5,240	40 mW
U-NII middle band (5.25–5.35 GHz)	52 56 60 64	5,260 5,280 5,300 5,320	200 mW
U-NII upper band (5.725–5.825 GHz)	149 153 157 161	5,745 5,765 5,785 5,805	80 mW

Figure 25.10 U.S. channel assignments for 802.11a.

In addition, four pilot subcarriers are used to transmit a BPSK-modulated pseudo-random binary sequence (PRBS), which is used by the receiver to detect frequency offsets and phase noise. The receiver can use the impairments detected in the known PRBS of the pilot carriers to compensate for impairments in the data subcarriers. All 52 carriers occupy a bandwidth of 16.6 MHz. In addition, convolutional coding is used to provide forward error correction at a coding rate of 1/2, 2/3, or 3/4.

The data subcarriers are modulated using BPSK, QPSK, QAM-16, or QAM-64 to achieve data rates of 6, 9, 12, 18, 24, 36, 48, and 54 Mbps, depending on modulation and FEC coding rate. However, only three data rates are mandatory in the specification: 6, 12, and 24 Mbps (corresponding to BPSK, QPSK, and QAM-16 at a coding rate of 1/2).

25.12 CableHome

CableHome is a relatively new initiative that focuses on the residential gateway and home networks. The CableHome 1.0 specification was released by CableLabs in January 2003.[22] The goals of the CableHome architecture are to simplify installation, to reduce equipment costs, and to reduce operational costs. CableHome specifies a home access (HA) device that performs the functions of a residential gateway and is the entry point into the home. The CableHome HA device may be manufactured as a stand-alone device or integrated with a DOCSIS cable modem. The intent of the specification is to allow multiple manufacturers to build HA devices that supports a wide variety of IP-based applications, including high-speed data, streaming audio and video, packetized telephony, network management, home security, environmental monitoring, medical monitoring, gaming, interactive television, and video conferencing.

It is important to note that the CableHome architecture does not attempt to reinvent any of the home network technology that has already been discussed in this chapter. Rather, CableHome aims to provide the cable operators with tools to better manage home networks deployed in their customers' homes by adding features to the residential gateway (or CableHome HA device).

The elements of the CableHome architecture are broken down into four main areas.

- *Management and provisioning* — CableHome defines a standard SNMP MIB for the HA device to allow secure SNMPv3 configuration and performance management by the cable operator. In addition, the HA device makes all "LAN IP devices" attached to the home network visible to a remote network management system.
- *Packet handling and address translation* — CableHome supports the use of network address port translation to allow many devices to share a single public

IP address provided by the cable operator. In addition, a nontranslated service is defined for applications that do not operate well behind NAPT.

- *Quality of service* — The HA device allows the various QoS messages to be exchanged with PacketCable (voice) applications so that QoS-protected flows can be established.
- *Security* — CableHome supports authentication of the HA device by the cable operator and allows secure management of the HA.

25.12.1 Management and Provisioning

One of the main goals of CableHome is to support remote management and provisioning of the home network from a network management system (NMS) that is part of the cable operator's back-office environment. This allows operations to be automated and can be used to enforce consistent practices across all home networks supported by the cable operator. CableHome defines an SNMP MIB for each of the various components of the HA, allowing an SNMP-based NMS to fully manage and provision them. CableHome includes support for secure software download (using TFTP) so that automated procedures can be established to maintain all HA devices at a consistent revision of software.

Figure 25.11 illustrates some of the key concepts in the CableHome model. The home access device contains an SNMP agent that receives commands and reports events to the NMS located at the cable operator headend. The SNMP agent not only interacts with the other components of the HA device but is also in a position to monitor the state of the home network and the LAN IP devices connected to it. For example, it can report address leases granted by its DHCP server and names assigned by its DNS server. However, because the LAN IP devices in the home

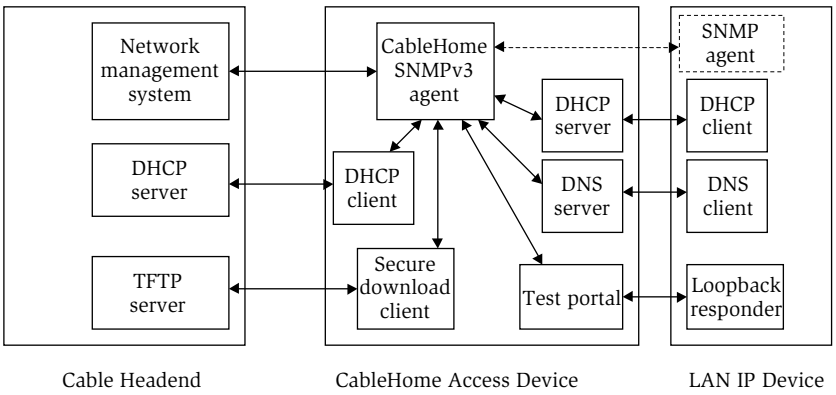


Figure 25.11 CableHome management and provisioning model.

are typically assigned nonroutable private IP addresses, the NMS has no direct visibility to them. For this reason, a test portal is defined that can ping the LAN IP devices under control of the NMS. The test portal can establish connectivity and speed of data transfer, providing information for remote diagnosis of a problem.

In addition, it is possible for the HA device SNMP agent to act as a proxy to allow access to an SNMP agent in the LAN IP device. (This flow is currently undefined in the CableHome specification.)

CableHome 1.0 defines the network management methods in six standard MIBS, allowing the cable operators to build a unified back-office system to manage CableHome devices from multiple vendors.

25.12.2 Packet Handling and Address Translation

CableHome follows the residential gateway model (see Section 25.6) for packet handling and address translation. Network address port translation allows the conservation of public IP addresses, as previously described in Section 25.6.2. However, because not all devices can operate behind an NAPT function, a nontranslated mode is also defined that can be configured for a specific port on the home network.

The configuration of all the address translation and packet-handling functions can be managed from an SNMP-based NMS, allowing the operator to create and administer standard configurations across all managed home networks.

25.12.3 Quality of Service

Quality-of-service (QoS) mechanisms can be used to provide prioritization of certain time-sensitive packets within the network. DOCSIS 1.1 supports special service flows designed to support multimedia traffic, such as voice over IP (VoIP). Ideally, QoS mechanisms should operate from end to end between the applications using it to be effective. However, the CableHome architecture does not specify how QoS can be extended from DOCSIS through the home network to the end application. To do this, a DOCSIS service flow must be mapped onto an equivalent flow over the home network; in most cases these do not yet exist. Nevertheless, excellent results can still be achieved if the home network segment operates at a high data rate compared to the application requirements and is not subject to traffic congestion.

PacketCable dynamic QoS service flows are established using messages that are exchanged between the CMTS and the message terminal adaptor (MTA). However, if the MTA is hidden behind a NAPT function, these messages will be blocked. To solve this problem, CableHome introduces a transparent bridge for PacketCable QoS messaging.

25.12.4 Security

CableHome defines a standards-based mechanism for authentication of the HA device using X.509 public key certificates and Kerberos. CableHome also supports secure

provisioning and management of the HA device. This is a considerable improvement over a stand-alone residential gateway, which cannot be authenticated or securely provisioned by the cable operator.

The LAN IP devices connected to the home network are protected from the outside environment using a firewall. Although CableHome provides a secure mechanism for loading a configuration file into the firewall, the actual configuration files themselves are specific to each manufacturer.

25.13 Summary

This chapter has provided a broad overview of the current state of home networking, focusing on the high-speed data applications. We have summarized the various applications and described the various home network technologies that have been developed to support them, including the residential gateway that connects a home network to a broadband connection. In particular, the CableHome initiative has focused on a specific form of a residential gateway that provides superior operational characteristics for the cable operator.

In the near future, the data rates supported by the various wired and wireless technologies will increase by an order of magnitude and support for quality-of-service guarantees will be added. These next-generation home networks will be able to support an integrated data, voice, and video environment, and we may see convergence between areas of the cable business that have traditionally been separate.

Endnotes

1. Ian Oliver and Brian Holmes, Installation and Ongoing Support Issues Relating to New Cable Services in the Home, *SCTE Emerging Technologies*, January 2003.
2. George Abe, *Residential Broadband*, 2nd ed. Indianapolis: Cisco Press, 2000, chap. 7.
3. IETF RFC 3022, *Traditional IP Network Address Translator (Traditional NAT)*, January 2001.
4. K. Sollings, IETF RFC 1350, *The TFTP Protocol (Revision 2)*, July 1992.
5. J. Case, R. Mundy, D. Partain, and B. Stewart, IETF RFC 3410, *Introduction and Applicability Statements for Internet Standard Management Framework*, December 2002.
6. Douglas Mauro and Kevin Schmidt, *Essential SNMP. Sebastopol, CA*: O'Reilly, 2001.
7. David Zeltserman, *Practical Guide to SNMPv3 and Network Management*.: Prentice Hall 1999.

8. *SNMP Link* Web site at http://www.snmplink.org.
9. William Cheswick and Steven Bellovin, *Firewalls and Internet Security — Repelling the Wily Hacker*. Reading, MA: Addison-Wesley, 1994.
10. Timothy Speed, Juanita Ellis, and Steffano Korper, *The Personal Internet Security Guidebook*. San Diego: Academic Press, 2002.
11. Stephen Palm, QoS Issues for In-Home Cable Services, *SCTE Emerging Technologies*, January 2003.
12. Paul Ferguson and Geoff Huston, *Quality of Service — Delivering QoS on the Internet and in Corporate Networks*. New York: Wiley, 1998.
13. Edward Frank and Jack Holloway, *Connecting the Home with a Phone Line Network Chip Set*, IEEE Micro, March–April 2000, available from www.homepna.org.
14. Mathew S. Gast, *802.11 Wireless Networks: The Definitive Guide*. Sebastopol, CA: O'Reilly, 2002.
15. Neil Reid and Ron Seide, *802.11 (WiFi) Networking Handbook*. New York: McGraw-Hill, 2003.
16. Richard Barrett, Wireless Connectivity Frees New Cable Applications, *SCTE Emerging Technologies*, January 2003.
17. ANSI/IEEE Std 802.11, 1999 Edition, *Information technology — Telecommunications and information exchange between systems — Local and metropolitan area networks — Specific requirements — Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*. Available at http://standards.ieee.org/getieee802.
18. The Wi-Fi Alliance Web page is www.weca.net.
19. IEEE Std 802.11b-1999 (supplement to ANSI/IEEE Std 802.11, 1999 Edition), *Higher-Speed Physical Layer Extension in the 2.4 GHz Band*. Available at http://standards.ieee.org/getieee802.
20. P. Nicopolitidis, M. S. Obaidat, G. I. Papadimitrion, and A. S. Pomportsis, *Wireless Networks*. Chichester, England: Wiley, 2003, chap. 9.
21. IEEE Std 802.11a — 1999 (supplement to ANSI/IEEE Std 802.11, 1999 Edition), *High-Speed Physical Layer in the 5-GHz Band*. Available at http://standards.ieee.org/getieee802.
22. CableLabs, *CableHome 1.0 Specification*, Revision I03, January 24, 2003. Available from www.cablelabs.com.

Appendix A

Channel Allocation

The channel allocation shown here conforms to the requirements of CEA/ANSI 542-B, which specifies the correspondence between channel number and frequency for North America. This standard was developed by the joint EIA (now CEA)/NCTA engineering committee, and the original version has been adopted by the FCC as the standard for channel allocations in the United States. (As of this writing the FCC Rules refer to the standard by its old working designation, IS-6.) Picture carrier frequencies are shown, even for QAM signals, which have but do not transmit, a carrier. The channel extends from 1.25 MHz below the analog picture carrier, to 4.75 MHz above. For QAM signals, the carrier is in the center of the channel. For 8-VSB signals, the carrier is 310 kHz above the low band edge. For all digital modulation, there is a single carrier, not separate carriers for picture and sound. Three channelization plans are shown. They are explained in section 9.2.

The complete standard is available from the Consumer Electronics Association's duplication contractor, Global Engineering, at 800-854-7179, http://global.ihs.com/. This is reprinted with permission of the Consumer Electronics Association.

	Channel Edge			Analog Picture Carrier			Digital Carrier					
Channel	Low (MHz) (Standard	High (MHz) and IRC)	Note	STD (MHz)	IRC (MHz)	HRC (MHz)	STD 8/16VSB (MHz)	STD QAM (MHz)	IRC 8/16VSB (MHz)	IRC QAM (MHz)	HRC 8/16VSB (MHz)	HRC QAM
1	72.0000	76.0000	+++		73.2625	72.0036			72.3100	75.0000	71.0600	73.7500
2	54.0000	60.0000		55.2500	55.2625	54.0027	54.3100	57.0000	54.3100	57.0000	53.0600	55.7500
3	60.0000	66.0000		61.2500	61.2625	60.0030	60.3100	63.0000	60.3100	63.0000	59.0600	61.7500
4	66.0000	72.0000		67.2500	67.2625	66.0033	66.3100	69.0000	66.3100	69.0000	65.0600	67.7500
5	76.0000	82.0000		77.2500	79.2625	78.0039	76.3100	79.0000	78.3100	81.0000	77.0600	79.7500
6	82.0000	88.0000		83.2500	85.2625	84.0042	82.3100	85.0000	84.3100	87.0000	83.0600	85.7500
7	174.0000	180.0000		175.2500	175.2625	174.0087	174.3100	177.0000	174.3100	177.0000	173.0600	175.7500
8	180.0000	186.0000		181.2500	181.2625	180.0090	180.3100	183.0000	180.3100	183.0000	179.0600	181.7500
9	186.0000	192.0000		187.2500	187.2625	186.0093	186.3100	189.0000	186.3100	189.0000	185.0600	187.7500
10	192.0000	198.0000		193.2500	193.2625	192.0096	192.3100	195.0000	192.3100	195.0000	191.0600	193.7500
11	198.0000	204.0000		199.2500	199.2625	198.0099	198.3100	201.0000	198.3100	201.0000	197.0600	199.7500
12	204.0000	210.0000		205.2500	205.2625	204.0102	204.3100	207.0000	204.3100	207.0000	203.0600	205.7500
13	210.0000	216.0000		211.2500	211.2625	210.0105	210.3100	213.0000	210.3100	213.0000	209.0600	211.7500
14	120.0000	126.0000		121.2625	121.2625	120.0060	120.3100	123.0000	120.3100	123.0000	119.0600	121.7500
15	126.0000	132.0000		127.2625	127.2625	126.0063	126.3100	129.0000	126.3100	129.0000	125.0600	127.7500
16	132.0000	138.0000		133.2625	133.2625	132.0066	132.3100	135.0000	132.3100	135.0000	131.0600	133.7500
17	138.0000	144.0000		139.2500	139.2625	138.0069	138.3100	141.0000	138.3100	141.0000	137.0600	139.7500
18	144.0000	150.0000		145.2500	145.2625	144.0072	144.3100	147.0000	144.3100	147.0000	143.0600	145.7500
19	150.0000	156.0000		151.2500	151.2625	150.0075	150.3100	153.0000	150.3100	153.0000	149.0600	151.7500
20	156.0000	162.0000		157.2500	157.2625	156.0078	156.3100	159.0000	156.3100	159.0000	155.0600	157.7500
21	162.0000	168.0000		163.2500	163.2625	162.0081	162.3100	165.0000	162.3100	165.0000	161.0600	163.7500
22	168.0000	174.0000		169.2500	169.2625	168.0084	168.3100	171.0000	168.3100	171.0000	167.0600	169.7500
23	216.0000	222.0000		217.2500	217.2625	216.0108	216.3100	219.0000	216.3100	219.0000	215.0600	217.7500
24	222.0000	228.0000		223.2500	223.2625	222.0111	222.3100	225.0000	222.3100	225.0000	221.0600	223.7500
25	228.0000	234.0000		229.2625	229.2625	228.0114	228.3100	231.0000	228.3100	231.0000	227.0600	229.7500
26	234.0000	240.0000		235.2625	235.2625	234.0117	234.3100	237.0000	234.3100	237.0000	233.0600	235.7500
27	240.0000	246.0000		241.2625	241.2625	240.0120	240.3100	243.0000	240.3100	243.0000	239.0600	241.7500
28	246.0000	252.0000		247.2625	247.2625	246.0123	246.3100	249.0000	246.3100	249.0000	245.0600	247.7500
29	252.0000	258.0000		253.2625	253.2625	252.0126	252.3100	255.0000	252.3100	255.0000	251.0600	253.7500
30	258.0000	264.0000		259.2625	259.2625	258.0129	258.3100	261.0000	258.3100	261.0000	257.0600	259.7500

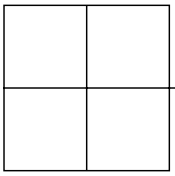
31	264.0000	270.0000	265.2625	265.2625	264.0132	264.3100	267.0000	264.3100	267.0000	263.0600	265.7500
32	270.0000	276.0000	271.2625	271.2625	270.0135	270.3100	273.0000	270.3100	273.0000	269.0600	271.7500
33	276.0000	282.0000	277.2625	277.2625	276.0138	276.3100	279.0000	276.3100	279.0000	275.0600	277.7500
34	282.0000	288.0000	283.2625	283.2625	282.0141	282.3100	285.0000	282.3100	285.0000	281.0600	283.7500
35	288.0000	294.0000	289.2625	289.2625	288.0144	288.3100	291.0000	288.3100	291.0000	287.0600	289.7500
36	294.0000	300.0000	295.2625	295.2625	294.0147	294.3100	297.0000	294.3100	297.0000	293.0600	295.7500
37	300.0000	306.0000	301.2625	301.2625	300.0150	300.3100	303.0000	300.3100	303.0000	299.0600	301.7500
38	306.0000	312.0000	307.2625	307.2625	306.0153	306.3100	309.0000	306.3100	309.0000	305.0600	307.7500
39	312.0000	318.0000	313.2625	313.2625	312.0156	312.3100	315.0000	312.3100	315.0000	311.0600	313.7500
40	318.0000	324.0000	319.2625	319.2625	318.0159	318.3100	321.0000	318.3100	321.0000	317.0600	319.7500
41	324.0000	330.0000	325.2625	325.2625	324.0162	324.3100	327.0000	324.3100	327.0000	323.0600	325.7500
42	330.0000	336.0000*	331.2750	331.2750	330.0165	330.3100	333.0000	330.3100	333.0000	329.0600	331.7500
43	336.0000	342.0000	337.2625	337.2625	336.0168	336.3100	339.0000	336.3100	339.0000	335.0600	337.7500
44	342.0000	348.0000	343.2625	343.2625	342.0171	342.3100	345.0000	342.3100	345.0000	341.0600	343.7500
45	348.0000	354.0000	349.2625	349.2625	348.0174	348.3100	351.0000	348.3100	351.0000	347.0600	349.7500
46	354.0000	360.0000	355.2625	355.2625	354.0177	354.3100	357.0000	354.3100	357.0000	353.0600	355.7500
47	360.0000	366.0000	361.2625	361.2625	360.0180	360.3100	363.0000	360.3100	363.0000	359.0600	361.7500
48	366.0000	372.0000	367.2625	367.2625	366.0183	366.3100	369.0000	366.3100	369.0000	365.0600	367.7500
49	372.0000	378.0000	373.2625	373.2625	372.0186	372.3100	375.0000	372.3100	375.0000	371.0600	373.7500
50	378.0000	384.0000	379.2625	379.2625	378.0189	378.3100	381.0000	378.3100	381.0000	377.0600	379.7500
51	384.0000	390.0000	385.2625	385.2625	384.0192	384.3100	387.0000	384.3100	387.0000	383.0600	385.7500
52	390.0000	396.0000	391.2625	391.2625	390.0195	390.3100	393.0000	390.3100	393.0000	389.0600	391.7500
53	396.0000	402.0000	397.2625	397.2625	396.0198	396.3100	399.0000	396.3100	399.0000	395.0600	397.7500
54	402.0000	408.0000	403.2500	403.2625	402.0201	402.3100	405.0000	402.3100	405.0000	401.0600	403.7500
55	408.0000	414.0000	409.2500	409.2625	408.0204	408.3100	411.0000	408.3100	411.0000	407.0600	409.7500
56	414.0000	420.0000	415.2500	415.2625	414.0207	414.3100	417.0000	414.3100	417.0000	413.0600	415.7500
57	420.0000	426.0000	421.2500	421.2625	420.0210	420.3100	423.0000	420.3100	423.0000	419.0600	421.7500
58	426.0000	432.0000	427.2500	427.2625	426.0213	426.3100	429.0000	426.3100	429.0000	425.0600	427.7500
59	432.0000	438.0000	433.2500	433.2625	432.0216	432.3100	435.0000	432.3100	435.0000	431.0600	433.7500
60	438.0000	444.0000	439.2500	439.2625	438.0219	438.3100	441.0000	438.3100	441.0000	437.0600	439.7500
61	444.0000	450.0000	445.2500	445.2625	444.0222	444.3100	447.0000	444.3100	447.0000	443.0600	445.7500
62	450.0000	456.0000	451.2500	451.2625	450.0225	450.3100	453.0000	450.3100	453.0000	449.0600	451.7500
63	456.0000	462.0000	457.2500	457.2625	456.0228	456.3100	459.0000	456.3100	459.0000	455.0600	457.7500
64	462.0000	468.0000	463.2500	463.2625	462.0231	462.3100	465.0000	462.3100	465.0000	461.0600	463.7500
65	468.0000	474.0000	469.2500	469.2625	468.0234	468.3100	471.0000	468.3100	471.0000	467.0600	469.7500

	Channel Edge			Analog Picture Carrier			Digital Carrier					
Channel	Low (MHz) (Standard and IRC)	High (MHz)	Note	STD (MHz)	IRC (MHz)	HRC (MHz)	STD 8/16VSB (MHz)	STD QAM (MHz)	IRC 8/16VSB (MHz)	IRC QAM (MHz)	HRC 8/16VSB (MHz)	HRC QAM
66	474.0000	480.0000		475.2500	475.2625	474.0237	474.3100	477.0000	474.3100	477.0000	473.0600	475.7500
67	480.0000	486.0000		481.2500	481.2625	480.0240	480.3100	483.0000	480.3100	483.0000	479.0600	481.7500
68	486.0000	492.0000		487.2500	487.2625	486.0243	486.3100	489.0000	486.3100	489.0000	485.0600	487.7500
69	492.0000	498.0000		493.2500	493.2625	492.0246	492.3100	495.0000	492.3100	495.0000	491.0600	493.7500
70	498.0000	504.0000		499.2500	499.2625	498.0249	498.3100	501.0000	498.3100	501.0000	497.0600	499.7500
71	504.0000	510.0000		505.2500	505.2625	504.0252	504.3100	507.0000	504.3100	507.0000	503.0600	505.7500
72	510.0000	516.0000		511.2500	511.2625	510.0255	510.3100	513.0000	510.3100	513.0000	509.0600	511.7500
73	516.0000	522.0000		517.2500	517.2625	516.0258	516.3100	519.0000	516.3100	519.0000	515.0600	517.7500
74	522.0000	528.0000		523.2500	523.2625	522.0261	522.3100	525.0000	522.3100	525.0000	521.0600	523.7500
75	528.0000	534.0000		529.2500	529.2625	528.0264	528.3100	531.0000	528.3100	531.0000	527.0600	529.7500
76	534.0000	540.0000		535.2500	535.2625	534.0267	534.3100	537.0000	534.3100	537.0000	533.0600	535.7500
77	540.0000	546.0000		541.2500	541.2625	540.0270	540.3100	543.0000	540.3100	543.0000	539.0600	541.7500
78	546.0000	552.0000		547.2500	547.2625	546.0273	546.3100	549.0000	546.3100	549.0000	545.0600	547.7500
79	552.0000	558.0000		553.2500	553.2625	552.0276	552.3100	555.0000	552.3100	555.0000	551.0600	553.7500
80	558.0000	564.0000		559.2500	559.2625	558.0279	558.3100	561.0000	558.3100	561.0000	557.0600	559.7500
81	564.0000	570.0000		565.2500	565.2625	564.0282	564.3100	567.0000	564.3100	567.0000	563.0600	565.7500
82	570.0000	576.0000		571.2500	571.2625	570.0285	570.3100	573.0000	570.3100	573.0000	569.0600	571.7500
83	576.0000	582.0000		577.2500	577.2625	576.0288	576.3100	579.0000	576.3100	579.0000	575.0600	577.7500
84	582.0000	588.0000		583.2500	583.2625	582.0291	582.3100	585.0000	582.3100	585.0000	581.0600	583.7500
85	588.0000	594.0000		589.2500	589.2625	588.0294	588.3100	591.0000	588.3100	591.0000	587.0600	589.7500
86	594.0000	600.0000		595.2500	595.2625	594.0297	594.3100	597.0000	594.3100	597.0000	593.0600	595.7500
87	600.0000	606.0000		601.2500	601.2625	600.0300	600.3100	603.0000	600.3100	603.0000	599.0600	601.7500
88	606.0000	612.0000		607.2500	607.2625	606.0303	606.3100	609.0000	606.3100	609.0000	605.0600	607.7500
89	612.0000	618.0000		613.2500	613.2625	612.0306	612.3100	615.0000	612.3100	615.0000	611.0600	613.7500
90	618.0000	624.0000		619.2500	619.2625	618.0309	618.3100	621.0000	618.3100	621.0000	617.0600	619.7500
91	624.0000	630.0000		625.2500	625.2625	624.0312	624.3100	627.0000	624.3100	627.0000	623.0600	625.7500
92	630.0000	636.0000		631.2500	631.2625	630.0315	630.3100	633.0000	630.3100	633.0000	629.0600	631.7500
93	636.0000	642.0000		637.2500	637.2625	636.0318	636.3100	639.0000	636.3100	639.0000	635.0600	637.7500
94	642.0000	648.0000		643.2500	643.2625	642.0321	642.3100	645.0000	642.3100	645.0000	641.0600	643.7500
95	90.0000	96.0000		91.2500	91.2625	90.0045	90.3100	93.0000	90.3100	93.0000	89.0600	91.7500

96	96.0000	102.0000	97.2500	97.2625	96.0048	96.3100	99.0000	96.3100	99.0000	95.0600	97.7500
97	102.0000	108.0000	103.2500	103.2625	102.0051	102.3100	105.0000	102.3100	105.0000	101.0600	103.7500
98	108.0000	114.0000*	109.2750	109.2750	108.0250	108.3100	111.0000	108.3100	111.0000	107.0600	109.7500
99	114.0000	120.0000*	115.2750	115.2750	114.0250	114.3100	117.0000	114.3100	117.0000	113.0600	115.7500
100	648.0000	654.0000	649.2500	649.2625	648.0324	648.3100	651.0000	648.3100	651.0000	647.0600	649.7500
101	654.0000	660.0000	655.2500	655.2625	654.0327	654.3100	657.0000	654.3100	657.0000	653.0600	655.7500
102	660.0000	666.0000	661.2500	661.2625	660.0330	660.3100	663.0000	660.3100	663.0000	659.0600	661.7500
103	666.0000	672.0000	667.2500	667.2625	666.0333	666.3100	669.0000	666.3100	669.0000	665.0600	667.7500
104	672.0000	678.0000	673.2500	673.2625	672.0336	672.3100	675.0000	672.3100	675.0000	671.0600	673.7500
105	678.0000	684.0000	679.2500	679.2625	678.0339	678.3100	681.0000	678.3100	681.0000	677.0600	679.7500
106	684.0000	690.0000	685.2500	685.2625	684.0342	684.3100	687.0000	684.3100	687.0000	683.0600	685.7500
107	690.0000	696.0000	691.2500	691.2625	690.0345	690.3100	693.0000	690.3100	693.0000	689.0600	691.7500
108	696.0000	702.0000	697.2500	697.2625	696.0348	696.3100	699.0000	696.3100	699.0000	695.0600	697.7500
109	702.0000	708.0000	703.2500	703.2625	702.0351	702.3100	705.0000	702.3100	705.0000	701.0600	703.7500
110	708.0000	714.0000	709.2500	709.2625	708.0354	708.3100	711.0000	708.3100	711.0000	707.0600	709.7500
111	714.0000	720.0000	715.2500	715.2625	714.0357	714.3100	717.0000	714.3100	717.0000	713.0600	715.7500
112	720.0000	726.0000	721.2500	721.2625	720.0360	720.3100	723.0000	720.3100	723.0000	719.0600	721.7500
113	726.0000	732.0000	727.2500	727.2625	726.0363	726.3100	729.0000	726.3100	729.0000	725.0600	727.7500
114	732.0000	738.0000	733.2500	733.2625	732.0366	732.3100	735.0000	732.3100	735.0000	731.0600	733.7500
115	738.0000	744.0000	739.2500	739.2625	738.0369	738.3100	741.0000	738.3100	741.0000	737.0600	739.7500
116	744.0000	750.0000	745.2500	745.2625	744.0372	744.3100	747.0000	744.3100	747.0000	743.0600	745.7500
117	750.0000	756.0000	751.2500	751.2625	750.0375	750.3100	753.0000	750.3100	753.0000	749.0600	751.7500
118	756.0000	762.0000	757.2500	757.2625	756.0378	756.3100	759.0000	756.3100	759.0000	755.0600	757.7500
119	762.0000	768.0000	763.2500	763.2625	762.0381	762.3100	765.0000	762.3100	765.0000	761.0600	763.7500
120	768.0000	774.0000	769.2500	769.2625	768.0384	768.3100	771.0000	768.3100	771.0000	767.0600	769.7500
121	774.0000	780.0000	775.2500	775.2625	774.0387	774.3100	777.0000	774.3100	777.0000	773.0600	775.7500
122	780.0000	786.0000	781.2500	781.2625	780.0390	780.3100	783.0000	780.3100	783.0000	779.0600	781.7500
123	786.0000	792.0000	787.2500	787.2625	786.0393	786.3100	789.0000	786.3100	789.0000	785.0600	787.7500
124	792.0000	798.0000	793.2500	793.2625	792.0396	792.3100	795.0000	792.3100	795.0000	791.0600	793.7500
125	798.0000	804.0000	799.2500	799.2625	798.0399	798.3100	801.0000	798.3100	801.0000	797.0600	799.7500
126	804.0000	810.0000	805.2500	805.2625	804.0402	804.3100	807.0000	804.3100	807.0000	803.0600	805.7500
127	810.0000	816.0000	811.2500	811.2625	810.0405	810.3100	813.0000	810.3100	813.0000	809.0600	811.7500
128	816.0000	822.0000	817.2500	817.2625	816.0408	816.3100	819.0000	816.3100	819.0000	815.0600	817.7500
129	822.0000	828.0000	823.2500	823.2625	822.0411	822.3100	825.0000	822.3100	825.0000	821.0600	823.7500
130	828.0000	834.0000	829.2500	829.2625	828.0414	828.3100	831.0000	828.3100	831.0000	827.0600	829.7500

	Channel Edge			Analog Picture Carrier			Digital Carrier					
Channel	Low (MHz) (Standard and IRC)	High (MHz) (Standard and IRC)	Note	STD (MHz)	IRC (MHz)	HRC (MHz)	STD 8/16VSB (MHz)	STD QAM (MHz)	IRC 8/16VSB (MHz)	IRC QAM (MHz)	HRC 8/16VSB (MHz)	HRC QAM
131	834.0000	840.0000		835.2500	835.2625	834.0417	834.3100	837.0000	834.3100	837.0000	833.0600	835.7500
132	840.0000	846.0000		841.2500	841.2625	840.0420	840.3100	843.0000	840.3100	843.0000	839.0600	841.7500
133	846.0000	852.0000		847.2500	847.2625	846.0423	846.3100	849.0000	846.3100	849.0000	845.0600	847.7500
134	852.0000	858.0000		853.2500	853.2625	852.0426	852.3100	855.0000	852.3100	855.0000	851.0600	853.7500
135	858.0000	864.0000		859.2500	859.2625	858.0429	858.3100	861.0000	858.3100	861.0000	857.0600	859.7500
136	864.0000	870.0000		865.2500	865.2625	864.0432	864.3100	867.0000	864.3100	867.0000	863.0600	865.7500
137	870.0000	876.0000		871.2500	871.2625	870.0435	870.3100	873.0000	870.3100	873.0000	869.0600	871.7500
138	876.0000	882.0000		877.2500	877.2625	876.0438	876.3100	879.0000	876.3100	879.0000	875.0600	877.7500
139	882.0000	888.0000		883.2500	883.2625	882.0441	882.3100	885.0000	882.3100	885.0000	881.0600	883.7500
140	888.0000	894.0000		889.2500	889.2625	888.0444	888.3100	891.0000	888.3100	891.0000	887.0600	889.7500
141	894.0000	900.0000		895.2500	895.2625	894.0447	894.3100	897.0000	894.3100	897.0000	893.0600	895.7500
142	900.0000	906.0000		901.2500	901.2625	900.0450	900.3100	903.0000	900.3100	903.0000	899.0600	901.7500
143	906.0000	912.0000		907.2500	907.2625	906.0453	906.3100	909.0000	906.3100	909.0000	905.0600	907.7500
144	912.0000	918.0000		913.2500	913.2625	912.0456	912.3100	915.0000	912.3100	915.0000	911.0600	913.7500
145	918.0000	924.0000		919.2500	919.2625	918.0459	918.3100	921.0000	918.3100	921.0000	917.0600	919.7500
146	924.0000	930.0000		925.2500	925.2625	924.0462	924.3100	927.0000	924.3100	927.0000	923.0600	925.7500
147	930.0000	936.0000		931.2500	931.2625	930.0465	930.3100	933.0000	930.3100	933.0000	929.0600	931.7500
148	936.0000	942.0000		937.2500	937.2625	936.0468	936.3100	939.0000	936.3100	939.0000	935.0600	937.7500
149	942.0000	948.0000		943.2500	943.2625	942.0471	942.3100	945.0000	942.3100	945.0000	941.0600	943.7500
150	948.0000	954.0000		949.2500	949.2625	948.0474	948.3100	951.0000	948.3100	951.0000	947.0600	949.7500
151	954.0000	960.0000	++	955.2500	955.2625	954.0477	954.3100	957.0000	954.3100	957.0000	953.0600	955.7500
152	960.0000	966.0000	++	961.2500	961.2625	960.0480	960.3100	963.0000	960.3100	963.0000	959.0600	961.7500
153	966.0000	972.0000	++	967.2500	967.2625	966.0483	966.3100	969.0000	966.3100	969.0000	965.0600	967.7500
154	972.0000	978.0000		973.2500	973.2625	972.0486	972.3100	975.0000	972.3100	975.0000	971.0600	973.7500
155	978.0000	984.0000		979.2500	979.2625	978.0489	978.3100	981.0000	978.3100	981.0000	977.0600	979.7500
156	984.0000	990.0000		985.2500	985.2625	984.0492	984.3100	987.0000	984.3100	987.0000	983.0600	985.7500
157	990.0000	996.0000		991.2500	991.2625	990.0495	990.3100	993.0000	990.3100	993.0000	989.0600	991.7500
158	996.0000	1002.0000		997.2500	997.2625	996.0498	996.3100	999.0000	996.3100	999.0000	995.0600	997.7500

++ These channels have been used as the 1ST IF in a few TVs. The use of these frequencies may result in interference on these TVs.
+++ Channel 1 is the gap between channels 4 and 5. It is only 4 MHz wide in the standard plan and is not used for television signals. In the HRC and IRC plants it opens to 6 MHz and may be used for television transmissions.
*Excluded from comb due to FCC offset in analog HRC.



Appendix B

Video Waveforms

This appendix illustrates the NTSC video waveform and is taken from EIA/TIA 250C, an earlier version of which formed the basis for the video waveform as specified in the FCC rules for broadcast television. See Chapter 2 for a more detailed description of the NTSC video waveform. The figures as shown in this appendix are drawn to conform as closely as possible to the defined amplitude and times of the NTSC signal. Other references frequently distort the time relationships.

The basic frequency in an NTSC waveform is the color subcarrier. All other video frequencies are defined in terms of the color subcarrier and for broadcast purposes are required to be locked to the subcarrier. The color subcarrier frequency is defined in the FCC rules as being

$$f_{\text{COLOR}} = \left(\frac{63}{88}\right) 5\,\text{MHz} = 3.5795454545\ldots\,\text{MHz}$$

The digits . . . 45 . . . repeat continuously. This is often called the 3.58 signal.

The horizontal line rate is

$$f_H = \frac{2}{455} f_{\text{COLOR}} = 15.734\,\text{kHz}$$

This frequency is sometimes referred to as 15.75 kHz, which was the old horizontal frequency prior to color transmission. At that time the horizontal frequency was exactly 15.75 kHz but was changed with the addition of color to reduce the effect of a beat between the color subcarrier and the sound carrier.

The time of one horizontal line is given by the reciprocal,

$$H = \frac{1}{15.734} = 63.5556\,\mu\text{s}$$

Since there are 525 horizontal lines in one frame and two fields in a frame, the vertical frequency is given by

$$f_V = \frac{2}{525} f_H = 59.94\,\text{Hz}$$

The resultant period of one field is then

$$V = \frac{1}{f_V} = 16.6833\,\text{ms}$$

From the above relationships, you can show that there are 227.5 cycles of the color subcarrier in one horizontal line, so when you trigger an oscilloscope to the horizontal sync, you will see an apparent phase reversal in the color burst from one line to the next. The color subcarrier is not really changing phase, but the extra half cycle per line results in the apparent phase reversal.

Figure B.1 shows the vertical interval of four consecutive fields. Together the four fields make up two frames, and it takes the two frames to return to the initial relation between sync and color subcarrier. Shown to the left of each line are the last active video lines at the bottom of the picture for each of four consecutive fields. The vertical interval begins with sync pulses spaced every half horizontal line rather than every one line. Each pulse is one half the normal width of 4.7 μs. These half-width, twice-occurring pulses are called *equalizing pulses*. All sync pulses are defined as starting at the negative-going edge of the pulse, which is the zero time reference for the blanking interval. If you start counting horizontal time intervals of 63.5556 μs from the negative-going edge of line 1, then all interval edges will fall on negative-going edges of horizontal sync pulses.

The start of odd fields is identified by the beginning of the equalizing pulses following a full width horizontal line. Even fields start with the first equalizing pulse following a half line. This is the way interlace is achieved. Notice that line 1 in even fields doesn't start immediately after the half active video line that ends the odd field.

Vertical sync is defined by three lines of video (lines 4–6), in which the voltage waveform is at the lowest level for all except 4.7μs in the center of the line and at the end of the line. Even in these three lines, the line begins at the negative-going transition of the waveform. Following common practice, the four vertical intervals are shown with the vertical sync pulses lined up. Vertical sync is followed by three more lines of equalizing pulses. During the nine line vertical interval, the color burst is not present. It resumes on line 10, occurring immediately after the sync pulse. Lines 10–20 are normal horizontal lines with sync and burst but no video. This is the *retrace interval*, which allows time for the electron beam to move from the bottom of the screen to the top. Video is never transmitted during these lines, but digital data and test signals may optionally be transmitted. Lines 17 and 18 are frequently used for vertical interval test signals (VITS), which may consist of various signals used to test the quality of the transmission path. Line 19 was at one time reserved for a vertical interval reference signal (VIRS), used to overcome some problems with processing signals in a network distribution system. Today, it may be used for a ghost-canceling reference signal.

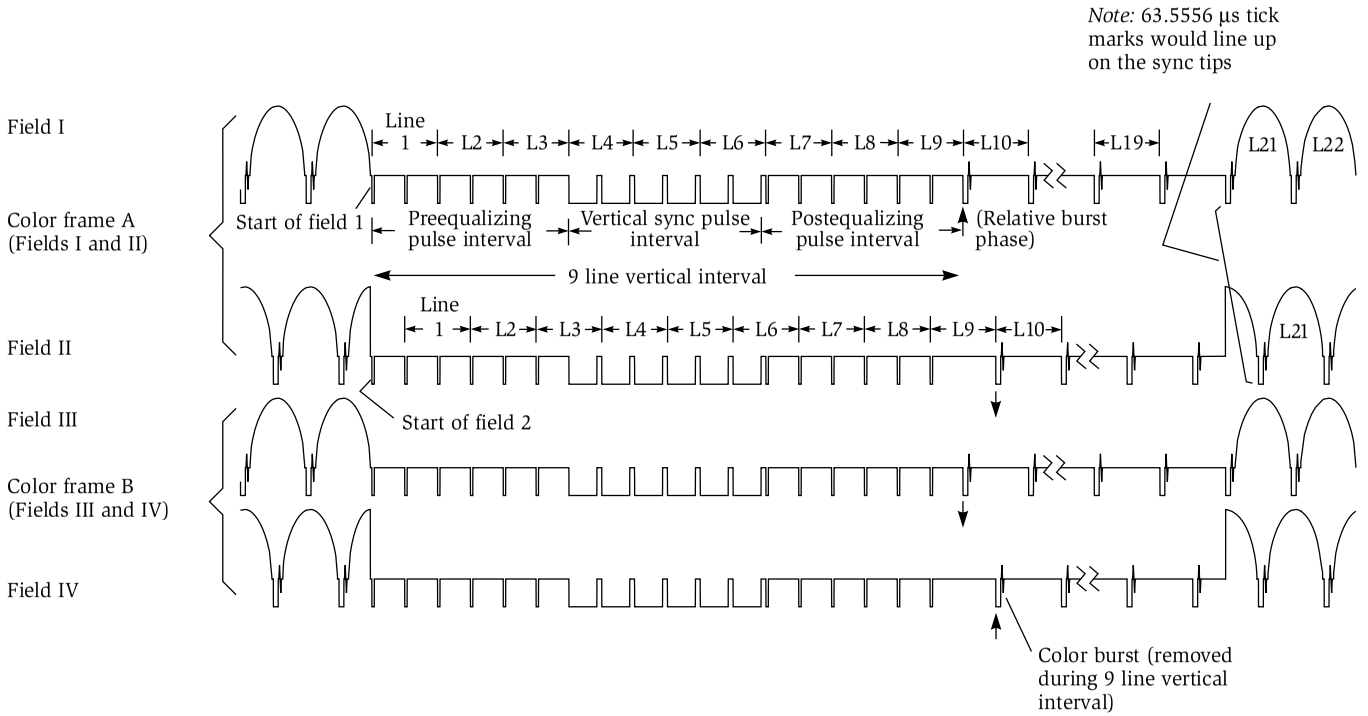


Figure B.1 Horizontal blanking interval.

Active video may start again on line 21, but most commonly today, line 21 is used for slow speed data transmission using a standard popularly known as the closed captioning standard. This is used for transmission of captioning data for viewers who are hard of hearing and also for transmission of a number of different types of information data for viewers.

Due to the way the color subcarrier is related to sync, the phase of the subcarrier relative to the start of sync changes between odd and even fields and from one odd (or even) field to the next odd (or even) field. This is indicated by the up and down arrows under the start of the sync pulse on line 10 in each field. This relationship is important when switching between two video sources. Fields I and III, the odd fields, are identical except that the phase of the color subcarrier is opposite in the two. Similarly, the even fields are identical except for the color phase.

Figure B.2 illustrates a typical horizontal blanking interval. All times are measured with respect to the 50% (−20 IRE) point of the negative-going edge of

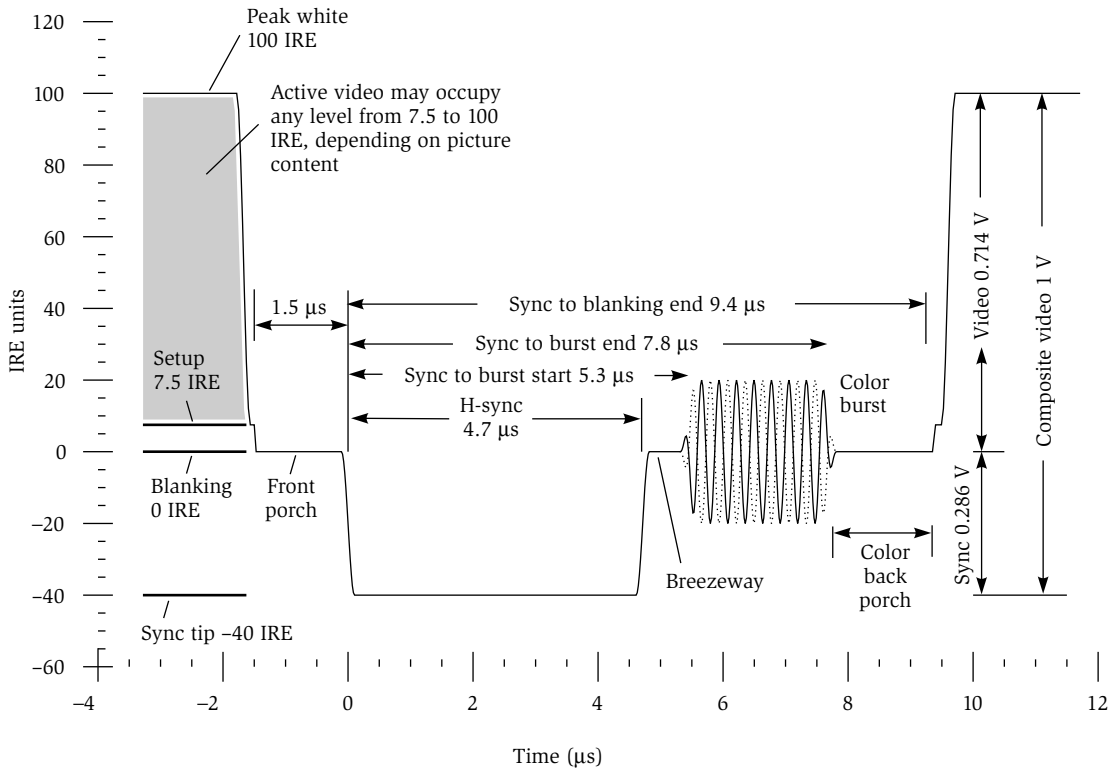


Figure B.2 Horizontal blanking interval.

the horizontal sync pulse. The video at the right side of the picture tube ends 1.5 μs before the start of sync. The period from the end of active video to start of sync is known as the *front porch*. It is during the front porch that most sync suppression scrambling systems switch to suppressing the sync pulse, creating an artifact as shown in Chapter 21.

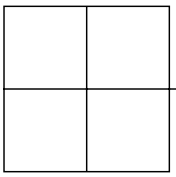
The actual sync pulse is 4.7 μs wide measured at the 50% (−20 IRE) points. The transitions from one level to another follow a sine squared shape with a duration of about 140 ns, limited by the 4.18-MHz baseband bandwidth of the NTSC signal. The sync is followed by a short *breezeway* between the end of sync and beginning of color burst. The color burst consists of 9 cycles of the 3.579-MHz color subcarrier. As explained above, due to the manner in which the sync relates to the color subcarrier, the subcarrier phase shifts with respect to the sync from line to line and field to field. It is shown here in both phases, because this is typically the way you will see the burst when observing it on a waveform monitor or oscilloscope. The color channel is limited in bandwidth, which is why the envelope of the color burst increases over about one cycle at the beginning of the burst and decreases over one cycle at the end. If the burst started and ended abruptly, the bandwidth of the signal would exceed the color channel bandwidth. The color burst is needed by the TV in order to demodulate the color subcarrier properly.

Following the color burst is another short interval before active video begins, called the *color back porch*. (In the black-and-white days, the color burst didn't exist, and the entire interval from sync tip end to active video begin was called the back porch. When the color burst was added, the name breezeway was used to define the interval between the sync and burst.) The color back porch is the region in which most sync suppression systems stop suppressing and expanding the sync, leaving another artifact. Normally the front and back porches are offscreen in the TV receiver, so the artifacts are not seen.

NTSC video levels are measured in IRE units, named for one of the predecessor organizations to today's IEEE, the Institute of Radio Engineers, which sponsored much of the early standardization of the NTSC system. The scale is referenced to the blanking level, occupied by the video waveform during the front and back porches. Sync extends from 0 IRE to −40 IRE. Active video occupies the positive portion of the IRE scale. It rides on a 7.5 IRE *setup* level, which in the early days was thought to improve the operation of low-cost tube sync separator circuits. Black corresponds to 7.5 IRE and white to 100 IRE, the maximum level of the video signal (sometimes the color subcarrier will extend somewhat above 100 IRE but not much above when the signal is transmitted). Video can begin at the same time setup begins, so the 7.5 IRE setup level may not be visible in the waveform.

The voltage amplitude of the composite (video and sync) signal is 1 volt peak to peak when the signal is interfaced from one piece of equipment to another. Sync is always negative-going. This means that the active video (including setup) occupies 0.714 volts and the sync tip 0.286 volts.

Traditionally video interfaces between two pieces of equipment are ac coupled. This means that the sync tip level will move up and down as the signal's average value changes with video content. 0 volts will be at the average value. Some modern systems clamp their output video so that 0 volts occur at the blanking level. The video is said to be dc coupled. Because each piece of equipment is capable of clamping the video as it requires, equipment with ac or dc coupling can usually be intermixed without regard for the coupling.



Glossary

Note that the terms contained herein are not intended to be rigorous definitions but rather to reflect common usage in the lexicon of the cable television industry.

absorption (optical) Attenuation of optical signals due to conversion of optical energy into heat energy.

actives Components that handle signals and that require power. Examples are line amplifiers and nodes.

adaptive equalizer A circuit (or software) that compensates for signal degradation in a path by adjusting taps in a delay circuit until the recovered waveform is as good as it can get.

addressability The ability to remotely control parameters of the subscriber's reception of signals by transmitting signals addressed to that subscriber's equipment.

AGC Automatic gain control. Circuitry in an amplifier that senses the level of a signal at the output and adjusts the gain such that the signal level is constant regardless of input level.

AGL Above ground level. Used in microwave systems to designate the elevation of various antennas relative to the ground level at the foot of the tower.

aloha protocol A data transmission protocol in which a device needing to transmit information on a common channel first listens on that channel and, if nothing is heard, transmits its information. It then waits for an acknowledgment from the intended receiver and retransmits after a random delay if the acknowledgement is not received.

alpha In data transmission, the excess width of a modulated RF signal compared with the minimum theoretical width needed to transmit data at a particular symbol rate.

AM/FM conversion AM-to-FM conversion. A type of distortion to a modulated signal in which amplitude modulation is translated into frequency modulation. Also known as AM/PM conversion.

AM/PM conversion AM-to-PM conversion. A type of distortion to a modulated signal in which amplitude changes in the signal are translated into phase changes. Also known as AM/FM conversion.

AML Amplitude modulated link. A microwave technology in which the cable television spectrum is up-converted to a microwave band for transmission and then down-converted at the receiver for transmission through a cable distribution system.

AMSL Above mean sea level. The official designation of elevation used in microwave systems.

analog television The system of television transmission that has been practiced since the medium was invented. Picture levels are represented by a continuously varying voltage level.

ANI Automatic number identification. A capability of most telephone systems that allows the number of the calling party to be known to the receiving party. Used in some pay-per-view ordering systems.

ANSI American National Standards Institute. An umbrella standards organization for the United States.

APC Angle polished connector. A fiber-optic connector that exhibits reduced reflections due to an angled fiber end.

application software Software that allows for any prescribed functionality in a set-top terminal.

artifact As used commonly in television technology, any artificial element introduced into a picture or sound signal as a result of the processing the signal undergoes.

ARU Audio response unit. A telephone answering system that guides the caller through audio menus and accepts inputs via Touch-Tone key presses. Used in some pay-per-view ordering systems.

ASC Automatic slope control, also known as dual-pilot AGC. A system in which one low-frequency carrier and one high-frequency carrier are both monitored for AGC purposes. By monitoring two carrier levels in different parts of the spectrum, the system can compensate for changes in gain over the entire spectrum. ASC systems tend not to be used in newer designs because of cost and the reduced need due to the relatively short cascades used with HFC architectures.

ASG Automatic slope and gain. A control system used in many modern pieces of distribution equipment in which an AGC function is combined with compensation for varying cable loss with frequency. As the gain at a pilot frequency is changed, the gain at higher frequencies is changed more and that at lower frequencies is changed less, matching an assumed profile for the coaxial cable.

ASI Asynchronous serial interface. A popular video interface for headend equipment. It uses 75-ohm coaxial cable or fiber-optic cable and operates at a wire rate of 270 Mb/s.

aspect ratio The ratio of width to height of a picture. Standard-definition television uses a 4:3 aspect ratio, and most HDTV systems use a 16:9 aspect ratio.

asymmetric key pair See *public/private key pair*.

ATM Asynchronous transfer mode. A set of standards for transmitting many types of baseband digital signals over a network. Very efficient for sharing channels among bitstreams of varying bit rates.

ATTC Advanced Television Test Center. An organization established to provide impartial testing for proponent digital video systems, which more recently has coordinated much of the testing required to make digital television, and especially HDTV, a reality.

attenuation A measure of the RF or optical signal loss (generally in decibels) of a component, circuit, or cable.

audio A baseband signal conveying at least one channel of sound information.

aural An adjective implying an RF or IF carrier that is modulated with an audio signal.

availability A measure of the fraction of time that a circuit or system is available to carry signals.

AVI Audio–video interface. A consumer interface for transferring digital video to a display.

azimuth The compass heading of an antenna, measured in degrees clockwise from north.

backbone Point-to-point or ring connections between headends or linking headends and major hubs in a cable system. Often signals traveling in the backbone are not in the same format used to deliver services to homes (for example, digital baseband).

balanced transmission A technique of coupling signals between two pieces of equipment using two conductors of equal and opposite signal voltage, neither of which is grounded. Used extensively for audio and infrequently for video and RF.

balun Balanced to unbalanced. A device that converts a transmission between balanced and unbalanced media.

bandstop filter A filter that passes all frequencies except those within some band. The term is specifically used to indicate a filter that goes in a drop, whose purpose is to prevent interference in the house from getting on the cable return plant, with the exception of a band of frequencies in the return spectrum that are allowed through, to support legacy services such as RF impulse pay-per-view.

bandwidth efficiency The ratio of bits per second transmitted through a communications channel to the channel bandwidth in hertz. The units are bits per second per hertz.

barker channel A channel used to promote programming on other channels. It is often automatically tuned when the subscriber attempts to tune a channel for which he or she is not authorized. It is implemented in some set-top converters.

baseband A signal not modulated onto an RF carrier. In video, refers to the visual signal as it is handled before being modulated onto an RF carrier.

basic service The foundation service all subscribers are required to take. Usually consists of at least local off-air channels and public channels and often is a reasonably complete set of signals.

baud In digital transmission, the number of symbols transmitted per second. Except for a few simple modulation formats, the baud rate is less than the bit rate.

Bellcore Bell Communications Research. The research and development arm of the Bell Telephone companies.

BER Bit error rate. A measure of the proportion of the bits in a data stream that are corrupted in transmission.

BERT Bit error rate test set. A test instrument used in data communications to insert a known data pattern at a transmitter and to count the number of errors received per second at a receiver.

binary Representation of a number by a series of elements (usually called "bits"), which can take on only one of two values. Contrast binary representation with our normal base-10 representation of numbers, where each element representing the number can take on any of 10 values.

bit A single binary data element having a value of either 1 or 0, usually represented by two different voltage levels.

bit rate The number of bits of data transmitted per second.

bit stuffing The process of adding extra data bits to a bitstream to get the data rate up to the transmission data rate. Practiced when the incoming data streams have a total bit rate less than that demanded by the channel.

blanking In a television picture tube, the process of temporarily shutting off the electron beam except during the "painting" of a picture, to avoid visible artifacts.

block conversion A technique for putting more signals on a return optical path by converting some incoming signals to alternate frequency bands.

BPSK Biphase shift keying. A digital modulation method in which the phase of a carrier is changed by 180° to represent transmission of a 1 or a 0.

bridger An amplifier (or module within a larger amplifier station) that is fed a sample of the trunk input signal and that is used to create a distribution leg.

broadband As used in the cable industry, refers to the ability to frequency division multiplex many signals in a wide bandwidth of RF frequencies, using an HFC network.

BTSC Broadcast Television System Committee. An entity that developed the analog TV stereo system used for transmission in North America and elsewhere.

byte A collection of eight bits.

CA certificate authority. An organization that can be trusted by both parties to a data exchange, which can sign a digital certificate proving the authenticity of the communicating parties.

CableLabs Cable Television Laboratories. The research consortium of the cable television operating companies.

caching A technique of retrieving information from Web sites and storing it on a local server, for the purpose of sending Web pages to users from that local server. The technique speeds up access to popular sites by eliminating the Internet delays involved in going back to the source for information. (The term also has an analogous meaning in retrieving data from memory.)

carrier Or RF carrier. A signal on which another, lower-frequency signal is modulated in order to transport the lower-frequency signal to another location.

carrier to noise (C/N) ratio In visual transmission, the ratio, usually expressed in decibels, of the RMS power present in the visual carrier during sync tips to the RMS noise power in 4 MHz. In cable television work, the term implies a measurement made at radio frequencies. Contrast with S/N, measured at baseband.

CARS Cable antenna relay service. The official FCC designation for a set of microwave frequencies reserved for the use of cable television companies.

cascade A configuration in which a signal passes through more than one stage or system. In cable usage, it usually refers to the system of line amplifiers through which the RF signals pass between the optical node and the subscriber. It can also refer to multiple stages of amplification or filtering through which a signal passes.

Cassegrain feed An antenna for microwave (including satellite) reception consisting of a parabolic main reflector that focuses power on a hyperbolic subreflector, which in turn focuses the power to the feed.

Cat 5 Category 5 cable. A data and voice transmission cable consisting of four pairs of twisted cables having a characteristic impedance of 100 ohms.

CBR Constant bit rate. An adjective describing a service or communications channel that carries a constant number of bits per second.

CCS 1. Custom calling services. A particular collection of enhanced telephone services, distinct from CLASS services. 2. Centum call seconds. Usage of a telephone resource (such as a phone line) for 100 seconds per hour.

CDMA Code division multiple access. A data transmission method in which a number of individual transmitters occupy the same spectrum at the same time, each using a different spreading sequence.

CEA Consumer Electronics Alliance. A branch of the Electronic Industries Alliance that represents manufacturers of consumer electronics equipment.

CEMA Consumer Electronics Manufacturers Association. A division of the Electronic Industries Alliance that represents manufacturers of consumer electronics. Name has been changed to CEA.

centum call second (CCS) In telephony, use of one telephone line for 100 seconds. There are 36 CCS in one hour per line.

channel A range of frequencies assigned to a signal in an FDM transmission system. In North America, analog video and most other downstream channels are 6 MHz wide and correspond to one of the allocation forms defined in the ANSI/EIA-542 standard.

channel mapping A feature of some set-top converters, TVs, and VCRs that allows programmable correspondence between channel number and received frequency. Useful where an off-air station must be channel shifted due to direct pickup problems. The station loses its channel identity to cable subscribers, but if channel mapping is available, it is possible to make the station appear on its off-air channel even though it is carried on another.

characteristic impedance The impedance of a transmission line such that, if terminated in a pure resistance of the same value, the return loss would be infinite. That is, the cable would "look" to a driving circuit like a resistor of the characteristic impedance.

chroma Or chrominance. Referring to the color information in a television signal, as opposed to the luminance (black and white) and aural (sound) portions.

chromatic dispersion The degree to which the transmission speed of an optical signal through a fiber or device varies as a function of wavelength.

chroma/luma delay A form of distortion of analog video signals in which the times of arrival of the luma (black-and-white) and chroma (color) signals are offset, resulting in the shift of color information with respect to the black-and-white outline. Also called chrominance delay.

CIN Composite intermodulation noise. Noiselike distortion products resulting from even- or odd-order mixing among RF signals, at least one of which is digitally modulated.

clamping See *dc restoration*.

cladding In optical fiber, the signals are transmitted primarily through the central, or *core*, glass, whereas the outer layers, or *cladding*, serve to add strength and, in combination with the core, to constrain the signals within the fiber.

CLASS Custom local area signaling services. A particular collection of enhanced telephone services.

CLEC Competitive local exchange carrier. A provider of telephone services who is not the original "incumbent" provider in a given market. Where cable television or OVS companies provide local phone service, it is almost always as a CLEC.

client A recipient of services provided by a central machine or process often called a server.

clipping A phenomenon where the voltage or current peaks of a signal are sharply limited at some value (for example, in directly modulated DFB laser diode transmitters).

closed captioning A system used to transmit captioning information to hard-of-hearing viewers, allowing text to overlay the picture. The system has been expanded to permit transmission of other types of data, most notably V-chip rating data.

CMI Cable microcell integrator. See *RAD*.

coax Short name for coaxial cable, which is used extensively for transmission of RF and baseband signals. The cable consists of an inner conductor on which signal voltage is impressed with respect to the shield. The center conductor is surrounded by a dielectric and then a shield. Frequently an insulation layer will surround the shield.

coaxial distribution The portion of the broadband cable television distribution network that starts at the node and terminates at the tap. Sometimes the drop system is included in the definition.

coding gain A measure of the efficiency with which a data stream is prepared for transmission. The ratio of the input to output bit rate of the encoder.

color difference signals The two signals in a TV picture that add color information. Derived from the original red, green, and blue primary signals, they are called color difference signals due to the way in which they are defined.

color subcarrier A signal in an analog TV signal that carries color information.

combiner A device with three or more ports used to combine two signals without regard for the frequency of the signals. Identical with a splitter, except for the direction of use.

community access Programming produced locally over which the cable operator has no editorial control. It may be produced by community groups, schools, etc.

composite video A baseband signal that contains multiple components of the signal required to reproduce a picture. The term has multiple meanings depending on context. It may mean a video signal including synchronizing (sync) information. It is also used to mean the signal including chroma (color) information.

compression A number of techniques for reducing the bandwidth needed to transmit a signal by removing unnecessary or redundant information from the signal.

concentration In telephony, the sharing of a resource among many users, based on the presumption that not all users will demand the resource at the same time.

conditional access A generic term used to indicate that access to a program or channel is limited to subscribers who have paid additional money for it.

cone of acceptance A theoretical cone whose point touches the end of a fiber whose axis is collinear with the fiber and that represents the maximum angles at which light may be introduced into the fiber and still have the transmitted portion of the light not "leak" out the sides due to refractive loss.

constellation diagram In data transmission, a diagram that plots the amplitude and phase of each state in a symbol.

CSO Composite second order. A form of distortion generated by amplifiers and other devices handling multiple RF signals. It is second-order distortion that combines signals at frequencies A and B, as $A \pm B$. The $A + B$ products fall 1.25 MHz above picture carriers, and the $A - B$ products fall 1.25 MHz below picture carriers in the standard channel allocation. The ratio of carrier to distortion product is often called C/CSO, carrier to composite second order.

CSR Customer service representative. An employee of a cable system who handles incoming telephone calls from customers.

CTB Composite triple beat. A form of distortion generated by amplifiers and other devices handling multiple RF signals. It is a third-order distortion product that combines signals A, B, and C as $A + B - C$. CTB products fall on the nominal frequency of carriers in the standard channel allocation. Also known as C/CTB.

cutoff frequency In filter terminology, a frequency at which the response of a filter is reduced by some specified amount, usually 3 dB.

CWDM Coarse wavelength division multiplexing. Wavelength multiplexing in which the combined wavelengths differ by 8–50 nm, but typically 20 nm.

data slicer In the recovery of a transmitted data signal, a circuit that compares the input signal voltage with a predetermined level. If the input signal is above the threshold, it is assigned one of the two logic levels. If it is below the threshold, it is assigned the other logic level.

dB Decibel. A measure of the relative strength of two signals. Defined as 10 times the logarithm (to the base 10) of the ratio of the two power levels, in any consistent units, such as watts, or 20 times the logarithm of the ratio of the voltage or current levels.

dBm Decibels with respect to one milliwatt. A unit of RF signal strength used in satellite work and other communications applications.

dBmV Decibels with respect to one millivolt in a 75-ohm system. The unit of RF power used in cable television work in North America. In a 75-ohm system, $0\,\text{dBmV} = -48.75\,\text{dBm}$.

DBS Direct broadcast satellite. A (primarily television) service in which signals are sent from a multichannel program supplier (for example, DirecTV or Dish Network) through a high-powered geostationary satellite to small-aperture receiving antennas and receivers typically mounted on each subscriber's house.

dc restoration In analog television transmission, the process of recovering a true black level in the picture by forcing the sync tip or back porch to be at a desired voltage. Also called clamping.

decibel See *dB*.

deemphasis In FM transmission, the process complementary to preemphasis. A frequency-selective reduction of the amplitude of higher modulating -frequencies after reception, to restore flat response. Used to improve the signal-to-noise ratio.

demodulation The process complementary to modulation. Recovery of information modulated onto a carrier.

demodulator A device used by cable television systems to convert an RF modulated video and/or audio signal to baseband.

DES Data encryption standard. A recognized data encryption system based on performing a prescribed set of operations on a block of plaintext data.

DFB laser Distributed feedback laser. A laser technology in which the wavelength of the laser is set by the distance between reflecting surfaces at each end of the laser cavity, augmented by a grating running the length of the cavity.

differential gain A measure of the variance in gain at the chrominance frequency as a function of luminance level.

differential phase A measure of the variance in phase at the chrominance frequency as a function of luminance level.

digital repeating ring An architectural structure in which a baseband digital signal is received and retransmitted at each terminal point around a closed ring.

digital return A return path from a node to a headend whereby the return signals are digitized and the digital signal is transmitted to the headend, where it is converted back to analog/RF form.

digital signature An electronic file that has been generated by a certificate authority (CA), which may be supplied from one member of a proposed data exchange session to the other member. The member receiving the signature is able to verify that it did, indeed, come from someone trustworthy.

digital television Television signals transmitted using digital techniques in which video levels are converted to digital states that approximate the analog level of the original signal. As used in the consumer world, it also implies compressing the signal to minimize the transmission bandwidth required.

digital video Representation of video signals in digital format. It does not necessarily imply a compressed video signal.

diplex filter A filter used to separate RF signals into one of two paths, based on the frequency of the signals. Can also be used to combine signals that exist in different frequency bands.

dipole A basic antenna consisting of a straight conductor usually broken in the middle to connect to a feed line.

direct pickup Or DPU. The characteristic of a TV or VCR by which a signal is picked up off the air in the device, creating interference with the signal being received via cable.

directional coupler A three port device for unequally dividing signal between two paths. Can also be used to combine signals with more loss in one input than in the other, with the signal from the high loss port being routed to only one of the other two ports.

directivity In a tap or directional coupler, the difference in sensitivity to upstream and downstream signals, as measured at the side port.

discrete cosine transform (DCT) A fundamental process used in video compression in which two-dimensional blocks of pixels are converted to a frequency domain representation, with the option of dropping higher-frequency components if they are small.

dispersion The variation of transmission time due to some controlling parameter.

dispersion-shifted fiber Most optical fiber exhibits a dispersion "zero" near the 1,310-nm wavelength. Dispersion-shifted fiber is modified so that the point of zero dispersion is moved to another wavelength, often near 1,550 nm.

distribution The broadband portion of a cable television network extending from the point where the broadband spectrum is created in the same format in which it will be received by subscribers (for example, in a simple star HFC network, the distribution system starts at the input to the linear optical transmitter).

DLC Digital loop carrier. A telephony system in which the analog interfaces of a switch are extended closer to the using telephones through use of a data stream to a remotely located interface.

DOCSIS Data-over-cable service interface specification. The formal name of the cable modem standard produced by a consortium led by CableLabs.

double sideband modulation Amplitude modulation in which both the upper and lower sidebands are retained. Also called DSB.

downstream Signal flow from a headend toward subscribers.

DPU Direct pickup. Signals picked up in the drop system and the customer's equipment due to inadequate shielding that may interfere with reception.

drop The portion of the distribution plant between the tap and the individual subscriber's home.

DSL Digital subscriber line. A method of sending digital information over existing twisted-pair telephone lines between a central office or remote terminal and an individual subscriber. Uses frequencies above the voice audio range and thus can operate simultaneously with normal analog baseband voice service.

DSM Demand side management. A method used by utility companies to limit peak usage of electrical power by controlling customer equipment.

DTCP Digital transmission content protection. The encryption system used with video transmission on IEEE 1394 networks.

DWDM Dense wavelength division multiplexing. Wavelength division multiplexing in which the optical channels are spaced 200 GHz or closer together.

EDFA Erbium-doped fiber amplifier. A common construction of optical amplifier, consisting of an erbium-doped fiber in which the signal to be amplified and a "pump" signal are inserted, with energy from the pump signal being converted into gain at the signal wavelength.

egress A measure of the degree to which signals from the nominally closed coaxial cable system are transmitted through the air. Also known as signal leakage.

EIA Electronic Industries Alliance. A trade organization in Washington, D.C., that represents companies that manufacture electronic equipment.

8B/10B encoding An encoding method used in gigabit Ethernet, ASI video interconnections, and other applications in which each 8-bit binary word is replaced by 10 bits such that a minimum number of data transitions per symbol is enforced.

EIRP Effective isotropic radiated power. The total power supplied to an antenna, multiplied by the directive gain of the antenna in a certain direction.

electronic program guide (EPG) A display on the TV of a selection of programs available to the subscriber. Generally, all the subscriber need do to see the program is to highlight it using cursor keys on a remote control transmitter and press a "select" button. A database cross-references the program identification to the channel and automatically tunes to it. The EPG may include other features, such as information on programs and the ability to program a VCR.

elevation 1. The number of feet that a location is above mean sea level (AMSL). 2. The number of degrees by which an antenna is tilted upwards from the horizontal.

emergency override A technique whereby, in an emergency, normal audio and/or video programming on all or most channels can be replaced with emergency announcements generated by an authorized entity.

encryption The process of rendering a digital signal unintelligible to any receiver that doesn't have some unique piece of information needed to recover that signal.

envelope In amplitude modulation, the shape that contains the amplitude modulated carrier. That is, the envelope is the outline of the carrier amplitude versus time.

envelope detection The process of recovering modulation from an amplitude modulated signal by recovering the envelope shape of the modulated signal. The most common envelope detector is a diode.

erlang In telephony, the maximum physical capacity of a line or trunk. One telephone line has a capacity of 1 erlang, or 36 CCS.

error correction In data transmission, the technique of adding extra bits to a transmitted signal, with the extra bits being used to detect and correct errors in the transmission.

Ethernet A popular layer 2 protocol used for interconnecting data equipment.

expanded basic service A subscription service offered by a cable company, which consists of basic service plus more advertiser-supported channels (as opposed to pay services).

express feeder An untapped coaxial line between two amplifiers in the coaxial portion of an HFC distribution system.

extinction ratio In a binary laser transmitter, the ratio of the maximum light output in one binary state to the minimum light output on the other binary state.

eye diagram In data transmission, a plot of the trajectory of either an in-phase or quadrature signal.

Fabry–Perot laser A laser technology in which the wavelength of the laser is set by the distance between reflecting surfaces at each end of the laser cavity.

failure rate The per-unit average rate at which a group of similar devices will randomly fail during their normal service life. Usually expressed in fraction or percentage per year.

FAT channel Forward application transport channel. The channel used to convey video and/or audio programming.

F connector The RF connector commonly used to attach drop cables to taps and other devices.

FCC Federal Communications Commission. The governmental body in the United States that oversees telecommunications entities. In other countries, the corresponding body may be called the Department of Communications or may be part of a Ministry of Posts, Telegraph, and Telecommunications.

FDM Frequency division multiplexing. The process whereby modulated signals occupying nonoverlapping spectra are combined for transmission through shared equipment. Used for transmitting multiple video and nonvideo signals to homes in conventional cable television networks.

feeder The portion of the coaxial subnetworks that starts at an amplifier fed by an express feeder or trunk and supplies signals directly to subscriber taps.

field One-half of a TV picture, consisting of alternate lines. Two fields make up a complete frame.

filtering In communications, usually refers to the transmission of power in some frequency range(s), combined with the nontransmission of power in other frequency range(s). In computer use, may also refer to the elimination of certain types of information, with passage of other types.

FireWire See *IEEE 1394*.

FITL Fiber in the loop. Used by the telephone industry to describe the deployment of fiber on the subscriber side of class 5 telephone switches. See *loop*.

flat loss The loss in a cable plant that is not a function of frequency. Typically, taps and attenuators have flat loss, as contrasted to coaxial cable, whose loss varies approximately as the square root of frequency.

FM/AM conversion FM-to-AM conversion. A type of distortion to a modulated signal in which frequency modulation is translated into amplitude modulation.

forward Downstream transmission of cable signals.

forward error correction (FEC) In data transmission, a process by which additional data is added that is derived from the payload by an assigned algorithm. It allows the receiver to determine if certain classes of errors have occurred in transmission and in some cases allows other classes of errors to be corrected.

Fourier series A mathematical expression relating the time domain representation of a periodic waveform with the frequency domain representation. It shows that any periodic waveform may be expressed as a series of harmonically related sine and cosine waves. (Sine and cosine waves have the same shape but are shifted in phase by 90°.)

frame One complete TV picture, consisting of two fields.

franchise An agreement with a local governing body that permits a cable operator to construct a cable system and that spells out the conditions that must be met by that system.

frequency agile The ability to operate on more than one frequency (but not simultaneously).

frequency response The gain as a function of frequency through a system or device, relative to the defined ideal. In the case of a cable distribution system, this is also known as peak-to-valley response, or P/V.

Fresnel zone In a microwave path, this describes a hypothetical ring centered on the path such that the distance from transmitter to ring to receiver is one-half wavelength longer than the distance from transmitter to receiver.

FSAN Full-service access network. A standard for passive optical networks based on ATM.

FSK Frequency shift keying. A low-cost, spectrum-inefficient, very robust form of digital modulation in which the frequency of a carrier is changed to represent data.

FTF Fiber-to-the-feeder. A distribution architecture in which fiber nodes feed coaxial feeder lines without any intervening coaxial trunk links.

FTTC Fiber-to-the-curb. A delivery technology in which signals are delivered from the headend to a small cluster of homes, completely in the optical domain. Most commonly the electronic signals at the end of the network are delivered over a short, completely passive network.

FTTH Fiber-to-the-home. A delivery technology in which signals are delivered from the headend to the home, completely in the optical domain.

ground block A passive device located on the subscriber's premises that allows a safety ground to be connected to the drop.

group delay Generally, the deviation from linear phase shift versus frequency through a circuit or network. This can occur due to reflections in a network. Also, it is a property of many practical filters. Signal power near the edge of the passband are delayed in traversing the filter with respect to signals away from the edge. This gives rise to waveform distortion, which can damage both analog and digital signals.

G/T "G over T." A figure of merit for a satellite earth station, expressing the ratio of the gain of the station to its effective noise temperature. Allows computation of the signal strength needed to receive a signal satisfactorily. The higher the station G/T, the lower the signal level required.

hard cable Sometimes called hard line. An informal term applied to the solid-jacketed coaxial cable used in the coaxial distribution portion of a cable system ahead of the subscriber tap. Also applied as an adjective to that portion of the distribution system.

harmony agreement An agreement among manufacturers of digital video scrambling systems whereby multiple addressable control systems can share a common scrambled video stream.

HDCP High-bandwidth digital copy protection. The encryption system used with video transmission on AVI video display interfaces.

HDT Host digital terminal. Digital headend equipment used to interface with a distribution plant.

HDTV High-definition television. Television with significantly more picture information (resolution) than that provided by a good NTSC or PAL television signal. Although the definition of HDTV is nonspecific, it usually implies about twice the resolution in both the horizontal and vertical directions and a wide aspect ratio.

headend The point at which all programming is collected and formatted for placement on the cable system.

header A set of bits added at the beginning of a transmitted data group that identifies the transmission and its source and destination.

HFC Hybrid fiber-coax. A network for transmitting signals modulated onto RF carriers that consists of RF carriers modulated onto optical carriers and transmitted on fiber-optic cable, followed by transmission of the RF carriers on coaxial cable medium. The majority of cable plant built since the early 1990s is in such an architecture.

high-pass filter A filter that passes all frequencies above its cutoff frequency but not below. The term is specifically used to indicate a filter that goes in a drop and whose purpose is to prevent interference in the house from getting on the cable return path.

horizontal blanking interval (HBI) That time in a television signal during which the electron beam in the receiver is moving from the right side of the screen back to the left to start a new line. Normally consists of a sync pulse and color burst.

host A generic term used in OpenCable work to define a device that is involved with receiving and decoding a digital video signal. Hosts are assumed to include either an interface for a POD module or the same functionality.

HRC Harmonically related coherent. A method of establishing picture carriers on a cable plant such that all carriers are harmonics of a single fundamental frequency close to 6 MHz.

hub A point in a distribution system where signals are converted to the form used to transmit them to subscribers. If a hub system is used, each hub typically will supply signals to a number of nodes.

Huffman coding A digital compression technique in which longer symbols (sequences of bits) that occur more commonly are replaced by shorter symbols and shorter less common sequences are replaced by longer symbols.

IEEE Institute of Electrical and Electronics Engineers. A professional organization of technologists interested in all aspects of electrical engineering.

IEEE 1394 A digital bus used to interconnect consumer audio and video equipment, among other things. Known as FireWire.

IIN Interferometric intensity noise. Noise in an optical link caused by mixing of direct and doubly reflected signals in the optical detector diode.

image response The response of a tuner to signals in a portion of the spectrum separated from the IF by the local oscillator frequency but on the wrong side of the IF.

incidental wavelength modulation A property of optical transmitters whereby the wavelength varies, along with the amplitude, as a function of the modulating signals, also known as chirp.

index of refraction The ratio of the speed at which light is transmitted in a vacuum to its speed in a material. It is a property of any material that can transmit light.

ingress Over-air signals that are inadvertently coupled into the nominally closed coaxial cable distribution system.

intercarrier Refers to a process in television reception in which the picture and sound carriers are mixed together to form a sound IF at the difference frequency (called the intercarrier frequency).

interdiction A program denial technique in which a jamming signal is inserted very near the picture carrier of programs for which the subscriber has not paid. The interdiction equipment is usually integrated with the tap. Signals are sent in the clear from the headend.

interlaced scanning The process of scanning every other horizontal line in a picture during one-half of the picture transmission, followed by scanning the intervening lines. The purpose is to reduce the flicker that is perceived by the viewer.

interleaving In data transmission, a technique by which the order of bits transmitted is changed. The bits are returned to the original order before error correction. If a burst of errors occurs, then the errored bits will be "spread out" before error correction is applied. This makes error correction more effective.

IP Internet Protocol. A particular layer 3 protocol that assigns addresses to different endpoints in a network and handles transport of data to and from those endpoints.

IPPV Impulse pay-per-view. Enhanced pay-per-view technology in which a subscriber can make a last-minute impulsive purchase of a program rather than having to order it in advance.

IPTV Internet Protocol television. A technology for sending video information over an IP network.

IR blaster A technique in which one device controls another through the controlled device's infrared (IR) input. For example, a set-top converter with electronic program guide may program a VCR using the VCR's IR input. Typically, the blaster will consist of an IR transmitter on the end of a cable, with the transmitter possibly taped to the device it is controlling, such that its transmissions can be picked up by the controlled device.

IRC Incrementally related coherent. A method of establishing picture carriers on a cable plant such that all carriers are spaced 6 MHz apart based on a master oscillator, but the carrier frequencies are offset 1.25 MHz from harmonics of 6 MHz.

IRD Integrated receiver-decoder. A device used in a cable television headend to receive a scrambled signal from a satellite and to convert it to baseband. Consists of a satellite signal receiver and a descrambler.

isolation In a tap or directional coupler, the difference between the amount of signal coupled to the side port and the directivity.

ISP Internet service provider. An organization whose business it is to connect users to the Internet.

IXC Interexchange carrier. An acronym for a long-distance telephone carrier.

JEC Joint Engineering Committee. A joint committee of the EIA-CEMA and the NCTA that met periodically to discuss technical issues of interest to both the cable television and the consumer electronics industries. It has been responsible for developing several standards in use by the two industries, such as the ANSI/EIA-542 band plan. The committee is inactive as this is written.

Kell factor The vertical resolution of an interlace scanned picture is less than the number of active lines in the picture, due to interlace. The Kell factor estimates the practical reduction and is commonly taken as 0.7.

key A piece of information used to encrypt or decrypt data by operating on the data with the key.

key exchange The process by which a data transmitter and receiver agree on the encryption key(s) to use in encrypting and decrypting data.

LAN Local area network. A group of devices in the same general location that share data.

laser In the context of cable systems, the device used to convert electrical signals to optical signals.

lashing The wire used to attach cable to a strand.

latency A net delay in the time information is delivered, from when it is transmitted. A satellite circuit introduces latency due to the propagation time to the satellite and back. A more common use of the term is with respect to video compression, where practical delays in the compression system delay the reception of the signal with respect to the origination.

LE Line extender. A single-input, single-output amplifier used to boost signals in a coaxial distribution plant.

LEC Local exchange carrier. An acronym used to identify a local telephone company.

local origination Programming over which the cable operator has editorial control. It is usually produced in the system.

local oscillator A circuit that generates a frequency that is mixed with a signal to produce a signal at a desired third frequency.

log periodic An antenna type consisting of multiple driven elements having a specified ratio of length and spacing and used to direct energy along a single axis. Performance is similar to that of a Yagi, except the bandwidth can be much larger.

loop Local office outside plant. Used to describe that portion of a telephone network extending from the local (class 5) switch to the network termination points at individual subscribers.

low-pass filter A filter that passes all frequencies below its cutoff frequency but none above.

luminance Or luma. Referring to the black-and-white portion of the television signal, as opposed to the chrominance (color) and aural (sound) portions.

Mach–Zehnder device A device used to modulate a light signal external to the source, resulting in less incidental waveguide modulation than with directly modulated sources.

Manchester encoding In data transmission, a method of embedding clock into data such that recovery is simple.

material dispersion A property of material that causes the speed of transmission of light waves to vary as a function of wavelength.

MCNS Multimedia Cable Network System. An industry consortium that produced the DOCSIS procurement standard for modems to be used on cable television networks.

media gateway A device that translates voice signals and sometimes signaling protocols between VoIP and switched voice systems.

MER Modulation error ratio. A measurement of the quality of a digital modulated signal that defines the average error between the proper location of all points in the constellation and their actual location.

mesh An architecture in which terminal points are connected to several other terminal points directly rather than just to a central point (a star) or the adjacent terminals in a ring.

messenger A support member in an electronic or fiber cable.

MGCP Multimedia Gateway Control Protocol. A popular thin client protocol for use with VoIP systems.

microreflections Small reflections occurring within both the coaxial distribution and drop portions of a cable plant. They can lead to impaired analog reception, group delay, and errors in digital signals.

middleware The layer of software that provides for standardized application software to run on any set-top terminal.

mixer A circuit that accepts two signals at different frequencies and produces a third frequency equal to the sum or the difference of the two input frequencies.

modal dispersion In a multimode fiber, the variation of transmission time among the various possible paths for the light.

modulation The process of imposing information on an RF carrier by varying some parameter of the carrier. The parameters that can be varied are the amplitude, the frequency, and the phase.

modulator A device used by cable television systems to convert a baseband video and/or audio signal from baseband to the RF channel on which it is to be transmitted.

MPEG Motion Picture Experts Group. A group that developed standards for digital compression of television pictures. The second generation of this standard, MPEG-2 is used, in one form or another, for transmission of digital television in all parts of the world that have introduced such transmission.

MPEG-2 The set of digital video decoding rules that form the basis of off-air and cable TV digital video transmission.

MPEG-4 An advanced set of digital video decoding rules. The basic MPEG-4 is intended to provide for very low data rate transmission of less than entertainment-grade video. MPEG-4 level 10 (also known as AVC, advanced video coding) is intended for entertainment-grade video. Compare with MPEG-2.

MSO Multiple system operator. A company that owns more than one separate cable system.

MTBF Mean time between failure. A measure of the average time interval between circuit restoration from a previous failure and the onset of the next failure in a network.

MTTR Mean time to restore. A measure of the average time required to restore proper network/circuit operation after a failure.

multimode An optical fiber that supports more than one transmission path and thus has a transmission delay whose uncertainty increases with length, thereby limiting bandwidth.

multipath A condition where an over-the-air signal may reach a receiver via more than one path from a transmitter, leading to degraded reception. Applies to all transmission systems and frequencies.

multiplex Use of a single data path for more than one set of data.

multiplexing The combining of many independent signals into one signal path for transmission. Cable systems operate in a *frequency division* multiplexing mode, in which many signals at different frequencies are combined in one signal path. Another common example is *time division* multiplexing, where many signals share the same path but are segregated in terms of time.

mux See *multiplex*.

NCS Network-based call signaling. A signaling profile of the MGCP specified for use with cable-based VoIP systems. Similar to TGCP.

NCTA National Cable Telecommunications Association. A trade organization that represents companies in the cable television industry.

negative trap A technology for denying analog television programs to subscribers who have not purchased them. Negative trap systems work by transmitting normal signals on the cable plant. At the tap of subscribers who have not bought the program, a trap is inserted to remove the picture carrier.

network-based PVR Broadcasts are saved on hard drive at the headend and played out to subscribers as a VOD program.

NICAM Near instantaneous amplitude companding and modulation. A system of transmitting stereo audio digitally in a TV signal, developed in the United Kingdom and used there and in a number of other countries.

NID Network interface device. One of many names for a box on the side of the house that contains any required interfaces between a telephone line or cable plant and inside wiring.

node 1. The portion of a coaxial cable plant served from one optical receiver. 2. The equipment that receives the linearly modulated FDM spectrum, converts it back to RF form, and transmits into one or more coaxial distribution legs.

noise Generally defined as undesired signals in a transmission channel. In television, the term is commonly applied to broadband thermal noise *excluding* discrete carriers and distortion products.

noise figure The ratio, expressed in decibels, of the noise produced by a device, over and above the thermal noise produced in a matched resistor at the input to that device. Higher-noise-figure devices require a higher signal level to deliver a given signal-to-noise ratio.

NPR Noise power ratio. A method of specifying the quality of a transmission path by loading it with noise, eliminating the noise in a narrow band, and measuring the amount of noise filled in by the transmission path.

NRZ Nonreturn to zero. A method of transmitting digital data bit by bit, with each bit immediately following the preceding one, with no time between.

NTSC National Television System Committee. An entity that developed the analog TV system used in North America and elsewhere. There were actually two NTSCs: the first developed the monochrome transmission system and the second added color.

NVOD Near-video-on-demand. A PPV service that offers sufficiently frequent start times for popular events (movies) such that it approaches on-demand availability. Half- to quarter-hour incremented start times are common.

Nyquist frequency In a sampled data system, the highest frequency that can be sampled and then recovered without distortion.

Nyquist slope filter A filter applied to a VSB modulated signal just prior to detection to restore flat frequency response.

OAM&P Operations, administration, maintenance, and provisioning. A generic term for software suites that allow centralized administering of a telephone network. A superset of what the cable industry knows as status monitoring.

OCAP OpenCable application platform. Standardized middleware that provides certain applications for a set-top terminal.

offset feed An antenna for microwave reception consisting of a parabolic reflector and feed, shaped such that the axis of the focus meets the reflector at or about at the edge of the reflector. Offset feeds are popular with very-small-aperture antennas because of reduced blockage.

OOB channel Out-of-band channel. A channel used to convey data to and from a set-top terminal other than the channel used to convey video and/or audio programming.

OpenCable A set of industry standards intended to define the next-generation digital consumer devices used by the cable TV industry, to encourage supplier competition, and to create a retail hardware platform.

optical modulation index (OMI) A measure of how much optical modulation is applied to a laser. Usually stated in terms of how much of the range from the laser's quiescent bias point to cutoff is taken up by the signal. The OMI may be expressed in terms of per-channel or composite signal.

orthogonal Literally, at right angles. Often used by the engineering community to describe two functions that are independent of each other.

oscilloscope A piece of test equipment that allows one to display the time value of a waveform.

overhead In data transmission, the portion of the data that is included to enable transmission. The remaining data transmitted is the payload.

overload In a receiver, the condition of supplying too much signal to that receiver, resulting in the generation of distortion in the receiver.

OVS Open video system. A regulatory alternate to conventional cable television franchising for the operator of a broadband video distribution network. OVS operators are regulated at both the federal and state or local levels. A greater portion of the bandwidth of an OVS system is available for lease to unaffiliated third parties than is the case for cable television franchisees.

PacketCable A set of interoperable interface specifications for delivering advanced, real-time multimedia services over two-way cable plant. PacketCable is built on top of DOCSIS.

packetized elementary stream (PES) In MPEG-2 transmission, a collection of bits containing data from one source only. May contain video, audio, or data.

PAL Phase alternate line. The television system used in most of Europe and many other places, in which the color carrier phase definition changes in alternate scan lines.

parity An extra bit added to a word to make the total number of logical 1s either odd (odd parity) or even (even parity). Used to check for errors that may have been introduced during data transmission or storage.

passband The frequency or wavelength region passed by a filter.

passive A component that handles signals but that requires no power of its own. Examples are taps and power inserters.

pay services Non-advertiser-supported channels for which a subscriber pays on a subscription basis. Also called premium services.

payload In data transmission, the portion of the data that is desired to be transmitted. Other transmitted data, such as error correction and headers, is considered overhead.

PCS Personal communications services. An alternative to cellular telephone service that uses a different frequency range and generally smaller cell sizes.

PDL Polarization-dependent loss. A property of a device or fiber that causes the loss to vary as a function of the polarization of the signal, where the polarization is measured normal to the transmission axis.

peak-to-average ratio In data transmission, the ratio, usually expressed in decibels, between the peak power of a signal and the average power.

PEG Public, educational, and government. A set of locally generated channels many systems are required to carry as a public service.

PES See *packetized elementary stream*.

phase noise A distortion in which a carrier signal is subjected to random variations in phase.

picture-in-picture A feature of some television sets and set-top converters that allows display of a small second picture within the main picture.

picture-to-sound ratio The ratio between the amplitude of the picture carrier and the associated sound carrier. Usually set to 15 dB in cable television work.

PID Packet Identifier. A header preceding a PES in an MPEG-2 transport stream that identifies the purpose of the PES. In a multiplex of several different TV programs, for example, each audio and video will contain its unique PID, which allows for the selection of the desired program out of the multiplex.

pilot A signal on the distribution plant that is used for AGC or ASG purposes. In most cases, the pilot is one of the normal analog picture carriers that has been so designated.

pixel Picture element. The smallest unit of a picture that can be represented as combinations of the three primary colors only, with no unique shape. Frequently it is assumed that a pixel has either a square or rectangular shape.

plaintext An unencrypted digital data stream that can be read by anyone who understands the data transmission format.

PMD Polarization mode dispersion. The property of a device or fiber that causes signals at one polarization to travel at a different rate than signals at another polarization, where the polarizations are measured normal to the transmission axis.

POD module Point of deployment module. A module intended to be plugged into a host, which includes the intelligence for descrambling a digital video signal and communicating with a headend for the purpose of controlling the delivery of scrambled digital video signals. The module is also called a cablecard.

POE Point of entry. Equipment that is placed between the drop and inside wiring of a dwelling and that processes some or all of the signals for specific services. One application is cable telephony, where the POE device interfaces standard analog in-home wiring to the RF distribution system. See *NID*.

polarization An expression of the orientation of the electric field in a radiated RF signal.

polarization diversity The use of the same frequency band to radiate two sets of signals by using orthogonal polarizations, usually either vertical and horizontal or left- and right-hand circular.

PON Passive optical network. An all-optical network that consists of nothing more than fiber-optic cable and splitters.

port A logical construct through which data is transferred from one process to another.

positive trap A technology for denying analog television programs to subscribers who have not purchased them. Positive trap systems work by adding a high-level interfering carrier between the picture and sound carriers. Subscribers who have paid for the program are supplied with a filter (trap) to remove the interfering signal.

POTS Plain old telephone service.

power inserter A passive device used to add ac power to the center conductor of a coaxial cable for the purpose of powering active devices.

power pack In a cable system, the module within an amplifier station (or other active equipment) that receives the power multiplexed with the RF signals on the distribution system and creates the voltages required by the station's internal circuits.

power supply In a cable system, the device that receives power from the utility, processes it for use in the cable system, and inserts it into the coaxial plant through a power inserter.

PPV Pay-per-view. A video service in which the operator transmits movies or events on a fixed schedule and customers have the option of purchasing an individual event or movie showing.

Predistortion A technique for reducing distortion in a linear transmission system by intentionally generating distortion at the transmit end, which distortion will be cancelled by the real distortion of the network.

preemphasis In FM transmission, a technique for reducing noise by selectively increasing the deviation of higher frequencies before transmission. Deemphasis is used to restore flat response at the receiver.

primary colors The three colors that can be combined in differing amounts to yield any color that can be perceived by the human visual system. For the emission spectrum used in television, the primary colors are red, green, and blue.

prime focus feed An antenna for microwave reception consisting of a parabolic reflector and a feed assembly (antenna and, usually, preamplifier and converter) located at the focus of the parabola.

processor Or signal processor. In cable television terminology, an equipment used in the headend to receive a television signal and reformat it for transmission on cable. Reformatting consists of amplification, bandpass filtering, shifting the relative levels of the picture and sound carriers, and, frequently, changing the carrier frequency (channel).

protocol In data transmission, a set of rules defining how various pieces of equipment or software will interact cooperatively.

PSIP Program and System Information Protocol. A series of tables transmitted as part of a digital television transport stream that gives the receiver information needed to recover a program.

PSK Phase shift keying. In data transmission, a modulation format that depends solely on differences in the phase of a modulated signal to convey information.

public/private key pair A set of keys used to encrypt and decrypt messages. One of the keys of the pair is used for encryption and the other for decryption. The key used for encryption cannot be used for decryption. Also known as asymmetric keys.

P/V Peak to valley. The peak-to-peak variation in the frequency response of a cable distribution network relative to the ideal response curve.

PVR Personal video recorder. A device that records and plays back a compressed digital video signal on a hard drive.

QAM Quadrature amplitude modulation. A digital modulation method in which the value of a symbol consisting of multiple bits is represented by amplitude and phase states of a carrier. Typical types of QAM include 16QAM (four bits per symbol), 32QAM (five bits), 64QAM (six bits), and 256QAM (eight bits).

QoS Quality of service. Techniques for ensuring the delivery of data in a reliable and timely manner.

QPSK Quadrature phase shift keying. A digital modulation method in which the state of a two-bit symbol is represented by one of four possible phase states.

quadrature At right angles. In transmission, often used to express independence between two signals.

quantization noise The noiselike uncertainty imparted to an analog signal when it is digitized. An alternate term is quantization error. The more bits used in an analog-to-digital conversion process, the smaller the quantization noise.

RAD Remote antenna driver. A technique by which signals from PCS antennas are frequency converted and transmitted through cable systems to remotely located base stations.

RAID drive Redundant array of independent (or inexpensive) disks. A category of disk drives that employ two or more drives in combination for fault tolerance and performance.

rain fade In a terrestrial or satellite microwave link, signal loss due to absorption and reflections caused by raindrops within the signal path.

reflectometer A device used to determine the location of a fault in a cable by propagating a signal down the cable and reading any signal reflected. May be used for RF or optical transmission.

refraction The process by which optical signals are transmitted through the interface between two dissimilar materials. See *index of refraction*.

remultiplex Or remux. The process of taking some signals from each of two or more channels and recombining them into one new channel.

renewability In video scrambling, the ability to change a scrambling system if it is broken, or the ability to refuse to communicate with a device that is found to be treating programming in ways not intended by the program owner.

resolution A measure of the "sharpness" of a picture. Usually measured in pixels in the horizontal direction and lines or pixels in the vertical. It is frequently the case that horizontal and vertical resolution are different.

retrace In a scanned television picture, the retrace is the process of moving the electron scanning beam back to the start of each line (horizontal retrace) or the top of the screen (vertical retrace). The beam is shut off ("blanked") during retrace to avoid picture artifacts.

return loss A measure of the amount of incident power that is reflected back toward the opposite end of a cable from a termination or other discontinuity in the characteristic impedance of the cable.

reverse See *upstream*.

RIN Relative intensity noise. A measure of the residual internal noise modulation of an optical transmitter. Often limits the attainable C/N of RF signals transmitted through an optical link.

ring network An architecture in which terminal points are connected in a continuous ring and each terminal is connected to both of its neighbors.

router In data communications, a device used to switch packets from any number of input ports to any one of a number of output ports.

RSA An encryption algorithm named for its developers: Rivest, Shamir, and Adleman.

RTP/RTCP Real-time protocol/real-time control protocol. Two layer-4 protocols used to transmit real-time information, and to control playout of packets at the proper time.

run-in A signal preceding some asynchronous data transmissions that permits the receiver to synchronize itself with the transmission.

SBS Stimulated Brillouin scattering. A mechanism that limits the amount of optical power at a single wavelength that can be transmitted through an optical fiber.

scanning The process of sequentially picking up, or "painting," a picture, wherein each picture element is delivered one at a time in a set sequence over the entire picture. Commonly, the scanning sequence is top to bottom and left to right, as viewed from in front of the picture tube.

scattering In an optical network, any process by which light is briefly delayed and then retransmitted. Forward scattering causes an overall slowing of the signal (see *index of refraction*), whereas scattering in other directions leads to signal loss and, sometimes, distortion).

scrambling In cable television, the process of rendering a video signal unintelligible without the knowledge of some secret information. In data transmission, scrambling refers to the process of randomizing the bit pattern of a transmitted signal to prevent peaks in the spectrum of the modulated signal. Note that the cable television and data transmission definitions are *not* equivalent.

SCTE Society of Cable Telecommunications Engineers. A professional organization of technologists interested in cable telecommunications technology.

SDM Space division multiplexing. The simultaneous use of different physical portions of a network (for example, nodes) to carry different signals at the same frequencies.

SECAM A television system used in France and some former Soviet bloc countries.

second audio program In BTSC stereo signals, a second, monaural, audio channel that is available to carry a second audio channel, such as a foreign-language soundtrack.

server A computer that sends information to client machines.

set-top converter Also known as a set-top terminal. A device used with a subscriber's TV to allow reception of television programs. It may tune channels the TV does not tune and may include descrambling circuitry. It also may include a digital decoder and auxiliary functions such as an electronic program guide.

sheath ring network A ring architecture in which separate fibers are used in each direction from a common terminal to each subterminal. Sheath rings require more fiber than repeating rings but are more secure and offer higher potential bandwidth to a common location.

shielding An expression of the ability of a device either to prevent radiation of signals contained within that device or to prevent undesired pickup of signals outside of the device.

shot noise In an optical link, noise at the optical receiver that is caused by the statistical variation in the arrival of photons.

signal level In analog television work, the RMS level of the picture carrier measured over the occupied bandwidth of the signal during sync tips. In digital communications, as practiced by the cable industry, it is the average level of the modulated signal or the level that would be read on a thermocouple-type RF power meter.

signal level meter A frequency selective RF voltmeter used to measure the level of individual RF carriers.

signal-to-noise (S/N) ratio In video work, the ratio between the video signal amplitude and the noise power measured in a reference bandwidth. Many variations in the definition of "signal" and "noise" exist, and the reader must be sure that the definitions in use are accurately conveyed.

single mode fiber Optical fibers that support only a single transmission path and thus that offer a better-defined transmission delay.

single sideband modulation Amplitude modulation in which one or the other of the modulation sidebands is removed.

skin effect The tendency of RF signals to travel near the surface of conductors.

SLM Signal level meter. A tunable power meter commonly used to measure signal levels in an FDM network.

softswitch A server in some VoIP systems that sets up and tears down calls by sending commands to the endpoints of the call in response to stimulus from the calling endpoint.

SONET Synchronous optical network. A set of standards describing an optical baseband network designed for carrying many types of data.

S/PDIF Sony/Philips Digital Interface. A standard digital audio interface.

spectrum analyzer A piece of test equipment used to plot amplitude vs. frequency.

splitter A device having three or more ports, used to divide RF signal equally between two paths, without regard for the frequency of the signal. Can also be used to combine two signals.

SPM Self-phase modulation. A distortion mechanism in optical fibers whereby, at high transmitted power levels, the optical signal instantaneously modulates the index of refraction of the glass and thus modulates the speed of transmission of the signal, ultimately leading to CSO distortion in the demodulated RF modulating signals.

spread spectrum Any of several techniques used to increase the transmitted spectrum of a signal with respect to the minimum spectrum that can be used for the transmission. The purposes are usually to reduce noise and interference pickup, to allow more than one signal to occupy the same spectrum, to mitigate transmission problems in one portion of the spectrum, or for security.

SRL Structural return loss. A measure of the degree to which signals are reflected from a length of cable as a function of frequency.

SRS Stimulated Raman scattering. A phenomenon in fiber whereby energy from one signal is transferred to one or more other signals sharing the fiber, leading to crosstalk among the RF signals after detection.

standard definition Refers to today's NTSC, PAL, and SECAM transmissions, with their normal resolution.

star network An architecture in which each terminal is directly connected through separate fibers to a common master terminal. In a simple HFC cable system, the master terminal is the headend and the other terminals are nodes.

start bit(s) One or more bits that precede a digital word in asynchronous transmission, used to identify the beginning of the transmission.

statistical multiplexing In multiplexed video compression systems, the process of assigning more or less of the available channel bandwidth to each of the several signals being multiplexed, according to the instantaneous needs of each.

stopband The frequency or wavelength region rejected by a filter.

strand A heavy wire attached to utility poles to which coaxial and fiber-optic cables are attached.

supertrunk In a cable system, the signal transportation link used to connect the point where the FDM signal complex is created to local coaxial distribution networks. In an HFC network, linearly modulated optical links are used, but other systems use AML microwave or coaxial supertrunks.

S-VHS Super-VHS. An analog video interface that carries the luminance and modulated chrominance signals on separate conductors. Also known as S-video.

S-VOD Subscription video on demand. A variation of VOD in which a subscriber pays a monthly fee to gain access to a set of VOD content.

switch In telephony, the device used to connect one telephone line to another for the duration of a call.

switched broadcast A technique whereby signals are delivered to the portions of a network serving groups of customers only when one or more customers request them. Unlike on-demand service, however, the signals are not spooled from a server, but only connected from real-time sources and thus are not under interactive control of customers.

switched circuit A data or voice transmission method in which resources are set up at the beginning of the transmission and removed at the end, with certain system resources being available continuously to the parties in communication during the call. Conventional telephony is switched circuit. Contrast with *VoIP*.

symbol A collection of some number of bits of data that are transmitted together.

symmetric key A key that may be used for encrypting and decrypting data, as opposed to an asymmetric key.

sync Or sync tip. The portion of the analog television signal used to synchronize scanning in the receiver to scanning in the camera. Corresponds to the most negative portion of the baseband waveform ("blacker than black"). In digital transmission, a set of bits of known value, used to locate a position in a data stream.

synchronization In data transmission, the process of ensuring that data recovery at the receiver is timed properly with the data being transmitted. In video, the process of ensuring that the camera pickup and the receiver display are scanning the same part of the image at the same time.

synchronous detection The process of recovering modulation from an amplitude modulated signal by mixing the signal with a replica of the original unmodulated carrier. In modulation methods with asymmetrical sidebands (such as analog television transmission), synchronous detection can be shown to provide more accurate detection than can the simpler envelope detection.

sync suppression scrambling A scrambling technique in which the synchronization portion of the picture is hidden, usually by moving it to a level normally occupied by picture information.

taboo channels In FCC assignment of off-air UHF channels, certain channels that cannot be used in a particular location due to the use of other channels. The taboo channels exist due to certain limitations in UHF converter performance that are brought about by the need to keep cost low.

tap The component installed in the distribution cable that diverts a portion of the downstream distribution signal and splits it to feed two to eight individual subscribers. In the reverse direction, it combines the signals from subscribers and inserts them into the upstream direction of the distribution cable.

TASO Television Allocations Study Organization. A group that in the late 1950s studied the subjective effects of noise and distortion on the television viewing experience. Used a six-point rating scale that is still referred to today.

TCP Transmission control protocol. A connection-oriented layer 4 protocol often used with IP to transmit data. Includes various mechanisms to enhance the probability that a packet will arrive.

TDM Time division multiplex. A data transmission method in which a number of individual digital data streams share a transmission channel, each occupying the channel for a portion of the total time.

TDMA Time division multiple access. A data transmission method in which a number of individual transmitters in different locations share a transmission channel, each occupying the channel for a portion of the total time.

TDR Time domain reflectometer. A device that measures the quality of a transmission line by transmitting a pulse and then measuring signals reflected as a function of time. The reflections indicate the distance to the point of reflection.

TEM Transverse electric and magnetic. A description of the normal field configuration for signals transmitted through coaxial cables.

10/100BaseT Two common forms of Ethernet transmission that transmit data at, respectively, 10 or 100 Mb/s.

termination The impedance at the end of a coaxial or fiber-optic cable that accepts power being transmitted on the cable.

TGCP Trunking Gateway Control Protocol. The call management interface protocol specified for use with cable-based VoIP systems. It is a profile of MGCP, similar to NCS.

thick client A term applied to a data communications user endpoint, such as a set-top terminal, that implies relatively high levels of embedded functionality, with the device depending more on itself and less on network resources for its functionality. Contrast with *thin client*.

thin client A term applied to a data communications user endpoint, such as a set-top terminal, that implies relatively low levels of embedded functionality, with the device depending more on network resources for its functionality. Contrast with *thick client*.

transition region The frequency or wavelength region(s) of a filter that are between the passband and the stopband(s).

transport stream In MPEG-2 systems, the data formatted for transport. Such data is composed of one or (usually) more PESs, plus overhead.

trap A passive coaxial component inserted in series with a drop line that blocks transmission of a specific frequency or range of frequencies. Used to block or enable specific channels or to tailor the services available in homes.

trunk 1. The untapped portion of a coaxial distribution network. Trunk links are typically operated at lower levels than distribution links, to minimize distortions. 2. An amplifier module used in the trunk portion of a coaxial distribution system. 3. An amplifier station that contains a trunk module. It may also contain a bridger module if it is desired to create a distribution leg at that location.

twisted pair A type of cable used extensively for transmission of balanced audio signals. The cable consists of two insulated conductors twisted on each other. It may or may not have a shield over the twisted pair.

UDP Universal Datagram Protocol. A connectionless layer 4 protocol often used with IP to transmit data. A simple protocol suitable for transmission of real-time audio and video information, for which retransmission of a lost packet is not feasible.

unavailability The fraction of time that a network or channel is unavailable for use. See *availability*.

universal remote control A remote control transmitter that can control more than one brand of TV, VCR, set-top converter, etc. Often used to control all of a consumer's entertainment equipment.

UPC Ultra-polished connector. An optical connector that minimizes loss and reflections by careful preparation and polishing of the fiber end and alignment with the mating fiber.

upstream Signal flow from subscribers toward a headend.

UTP Unshielded twisted pair. A data cable consisting of one or more twisted pair without shielding. See *Cat 5*.

VBI Vertical blanking interval. That time in a television signal during which the electron beam in the receiver is moving from the bottom of the screen back to the top in order to start a new field. The VBI is 11 horizontal lines long, and these lines are frequently used to insert test signals and data.

VBR Variable bit rate. A signal or communications channel that generates or accommodates a digital signal whose bit rate varies with time.

video In television work, this word is reserved to mean a baseband electrical signal that conveys information regarding picture content.

video compression The reduction of bandwidth needed to transmit or store a video signal by removing information in the signal that is not necessary for satisfactory perception of the signal or that can be inferred from other elements of the transmitted signal.

video inversion A scrambling technique in which analog video is inverted so that black corresponds to the minimum carrier amplitude, as opposed to the maximum amplitude (in NTSC and PAL transmission). Sync may or may not be inverted.

visual An adjective implying an RF or IF carrier that is modulated with a video signal.

VITS Vertical interval test signals. Signals that are inserted into the VBI to allow testing of the quality of the transmission link independent of the video programming.

VOD Video on demand. A video service that allows users to select the program and exact start time interactively. In some embodiments, it allows VCR-like control of the playback (for example, pause, rewind, fast-forward).

VoIP Voice over Internet Protocol. A methodology for transmitting digitized voice telephone signals by converting them into packets that can be transmitted through a standard Internet Protocol network that may be shared with other services.

VSB Vestigial sideband. The elimination of a large portion, but not all, of one sideband of an amplitude modulated signal. This may be practiced with either analog or digital modulation.

water peak The property of fiber that causes it to exhibit a loss peak near 1,383 nm due to the presence of trace amounts of hydroxyl ions. Some modern fibers no longer exhibit this phenomenon.

waveguide dispersion The property of fiber that causes the velocity of propagation of an optical signal to vary as a function of wavelength due to the changing mode field diameter and, thus, the relative influence of the differing indices of refraction of the core and cladding glass.

wavelength The distance traveled through a transmission medium by an RF or optical signal in a time equal to the inverse of the frequency (for example, for a 1-GHz signal, the distance traveled in 1 ns). It is equal to the speed of propagation in the medium, divided by the frequency.

wavelength dispersion In an optical fiber, the variation in transmission time as a function of the wavelength of the light.

wavelength multiplexer A device for combining or separating signals of differing wavelengths.

W curve A description of the variation of subscriber perceptibility of interfering signals in an analog television channel as a function of frequency within the channel boundaries. So called because of the shape of the curve, caused by the extreme sensitivity to interfering signals occurring near the visual and chrominance carriers.

WDM Wavelength division multiplexing. The use of an optical fiber to simultaneously carry signals at different wavelengths.

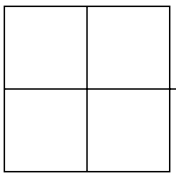
wire rate The actual data rate at which a signal is transferred on an interface, including all overhead.

WWDM Wide wavelength division multiplexing. Wavelength division multiplexing in which the optical signals are greater than 50 nm apart. The combination of a 1,310-nm and a 1,550-nm signals, for instance, would be considered WWDM.

XMOD Cross modulation. A type of distortion in which the amplitude modulation of one signal causes incidental amplitude modulation to another signal. XMOD is an odd order distortion.

XPM Cross-phase modulation. An interference phenomenon in optical fiber whereby a high-level signal at one wavelength modulates the index of refraction of the fiber and, thus, the speed of transmission of one or more other signals. When combined with wavelength dispersion, this leads to crosstalk in the detected RF signals.

Yagi Or Yagi-Uda. A gain antenna consisting of a single driven element and one or more parasitic elements used to direct energy along a single axis.



Index

ABC, 59
Above ground level (AGL), 601
Above mean sea level (AMSL), 601
Absorption, light, 514
A/B switch isolation, 948
Abramson, Norman, 209
Adaptive equalization, 165–168
Adaptive filter, 198
Addressability, 20–21
  in-band, 838
  out-of-band, 837
Addressing
  Ethernet, 216
  Internet Protocol, 221
  residential gateway and, 960, 961–963
Advanced Television Systems Committee (ATSC), 893
  Digital Television Standard, 111, 112, 129
  DTV Applications Software Environment (DASE), 894
  Program and System Information Protocol, 117–118
Advanced time division multiple access (ATDMA), 198
Advertising, inserting, 385–391
Aeronautical band plans, 400–402
A-law encoding, 247, 273
Aloha protocols, 189–190, 209–210
American National Standards Institute (ANSI), 129–130, 194, 847, 848, 884, 936
American Television and Communications, 929
Ampere's law, 244
Amplifiers
  characteristics, 684
  design choices, 462–465
  distortion, 442–456
  distribution, 463
  erbium-doped fiber, 540–542
  error, 469
  failure rates for, 804
  feed forward, 469
  fiber optics, 540–542
  frequency response variation, 461–462
  gain-flattened, 563–564
  group delay, 456–457
  hum modulation, 459–461
  impedance matching, 462
  line extenders, 463
  noise, 441–442
  operating dynamics, 465–468
  parallel hybrid, 469
  push-pull, 469
  random noise, 440–441
  stability, 458–459
  technology choices, 468–469
  transimpedance, 543, 771
  trunk, 463
Amplitude modulated link (AML), 18
  See also Microwave signals
Amplitude modulation
  mathematical approach to, 42–43
  phenomenological approach to, 40–41
Amplitude ripple, 165
AM sublow band, 328–329
AM supertrunk, 328
Analog copy protection, 864
Analog interfaces, 861–865
Analog modulation
  combining digital and, 772–773
  difference between digital and, 768–769
  receivers, 771–772
  transmitters, 769–771
Analog shared ring architecture, 744–745
Analog television
  ancillary signals, 58–63
  baseband monochrome video and monaural audio, 27–40
  color, 47–56
  compatibility, 27–28, 47, 49
  defined, 28
  modern, 47–63
  modulation, 40–46
  scanned images, spectrum of, 37–39
  scanning, 28–34
  signal quality, 63–66
  stereo sound, 57–58
  synchronization, 34–37
  television audio, 39–40
Analog-to-digital conversion, 245–247, 672
Ancillary signals, 58–63
Angle-polished connectors, 529–530
Antennas
  Cassegrain reflector, 308–309
  converter switch feature, 906
  dipole, 288–293
  horn, 311
  interference and ghost reduction, 296–300
  log periodic, 295–296

Antennas (*continued*)
multiple feeds, 310–311
for off-air reception, 288–300
offset feeds, 309–310
parabola, 305–307
prime focus feed, 307–308
rabbit ears, 299
satellite, 304–311
signal propagation and effects of height of, 301–302
ultra directive, 311
Yagi, 293–295
Antiope, 60, 61
Aperture mask, 47
Application layer, OSI, 188
Architecture(s)
broadcast video services requirements, 718–721
cable area network, 751–752
centralized, 759–765
changes in high-level, 19–20
distributed, 756–758
DOCSIS reference, 194
fiber deep, 775–786
fiber-to-the-curb, 524, 735, 752, 775–776
fiber-to-the-home, 776–778
information capacity, 702–712
multilayer, 754–756
narrowcast video services requirements, 721–723
on-demand/interactive video services requirements, 723–724
packet data services/Internet access requirements, 724–727
performance parameters, 701–718
reliability and availability, 712–716
ring, analog shared, 744–745
ring, digital repeating, 745–747
sheath ring, 743–744
simple passive, 497–502
star, 742–743
telephone digital switch, 233–235
telephone flat, 232–233
voice-grade telephony service requirements, 727–730
Architecture, elements of coaxial distribution network, 736–739
fiber interconnects, 741–747
fiber nodes, 739–741
switched video headend structures, 747–750
terminal equipment, 733–736
Architecture, examples of basic requirements, 750–751
centralized, 759–765
distributed, 756–758
historic development and small systems, 751–752
multilayer networks, 754–756
node size, 752–754
Aspect ratio, 45
Asynchronous digital hierarchy, 247–249
Asynchronous serial interface (ASI), 376–378, 385
Asynchronous Transfer Mode (ATM), 112, 208–209
AT&T, 60, 282
Digital Data System (DDS), 727
ATSC. *See* Advanced Television Systems Committee
Attack of the killer bees, 101
Attenuation
in bent fibers, 523–524
function of characteristic impedance, 430–432
function of frequency, 427–430
function of temperature, 430
interference mitigation and upstream, 687–689
selection based on maps, 689
selection based on tap value, 688
single-mode fiber optics and, 519–521, 523–524
Attenuators
midspan, 689
optical, 531
Audio
analog television monaural, 27–40
compression, 108–112
encoding and decoding, 109
masking, 108
monitoring, 419–420
scrambling, 836
wiring, 381–383
Audio distribution amplifiers (ADAs), 381
August mean radio refractivity, 602
Aural demodulation, 360–362
Aural modulation
audio monitoring, 419–421
diagram of, 346
limitations, 347
preemphasis, 345–347
process, 347–348
sound carrier deviation, 417–418
Authentication and key exchange (AKE), 890–891
Authentication certificates, 203
Auto bypass switch, 905
Auto channel setup, 906
Automatic gain circuit (AGC)
upstream, 659–660
video switch and, 342
Automatic level control (ACL), 658–660
Availability. *See* Service reliability and availability
Azimuth
for microwave paths, 599
for satellites, 311–312

Backfeed, 947, 948
Backup power
cable telephony, 253
telephone, 238
Baird, John Logie, 29
Baluns, 292–293, 294–295

Bandpass filters (BPFs), frequency shift keying and, 139
Band plans
aeronautical, 400–402
cable television, 393–402
EIA-542, 397–404
FM, 402
harmonically related coherent plan, 400
incrementally related coherent, 399–400
off-air, 398–399
Bandwidth
allocation, 199
coaxial network, 737
downstream 702–703
efficiency, 708–709
noise, 176
Nyquist, 245
per customer, 704–707
per-user, 13
signals and required, 156–157
total system, 13
upstream, 703–704
video, 43, 45
VoIP signals and, requirements, 280–281
Barker channel, 924
Baseband spectrum, of data signals, 141–142
Baseband transmission, 327–328
Base PID, 118
Battery, 240, 253
Baud rate, 154–155
BB audio interface, 864–865
BB video interface, 861
Beamwidth, antenna, 296
BellCore, 13, 243, 728, 729, 792, 809
Bessel null technique, 417–418
B frames, 98
Bidirectional transmission, 774–775
Biphase mark, 141
Biphase shift keying (BPSK), 142–145
bandwidth, required, 156
Bit error rate (BER), 142
measuring, 177–178
to set operating level, 667–668
versus carrier-to-noise ratio, 169–170
Bit rate, 154–155
constant, 389–391
variable, 389–391
Bits, 137
Blanking, 32
pulse, 35
Block code, 127
Block conversion, 653–655, 705
Block formation, 84, 102
Blocking noise, 100
Blocking of channels, 20
British Broadcasting Corp. (BBC), 60, 368
Broadband receivers, 941–942
Broadcast digital television, 72–78
Broadcast digital video services, 716
Broadcast receivers, 941–942
Broadcast Television System Committee (BTSC) stereo, 836
generation and application, 368–370
interfacing and setting levels, 370–371
Broadcast video services requirements, 718–721
Broad pulses, 35
Bucking array, 299
Burst protection period, 180

Cable Act (1992), 903, 923, 925, 930–933
Cable Act (1996), 7
Cable antenna relay service (CARS), 18, 935–936
reserved frequencies for, 594
Cable area network (CAN), 751–752
Cable drop. *See* Drop cable
CableHome, 979–982
CableLabs (Cable Television Laboratories), 8, 15, 194
CableHome, 979–982
CW tester, 677
direct pickup interference, 944, 945
early testing, 63–65
failure rates, 803–807, 809
OpenCable, 130–131, 848, 894–900, 937
Outage Reduction, 719–720, 790–792
PacketCable, 272, 273, 276, 848–849
standards, 847, 848, 937, 951
Cable modems, DOCSIS protocol for, 194–207
Cable modem termination system (CMTS), 167–168, 188, 195, 203
Cable network design, 14–16
Cable Network Interface specification, 132
Cable services
categories of, 903
interfaces and, 903–905
Cable telephony
See also Telephone network digital hierarchy; Telephone systems, architecture
channel sharing, 255–256
DS1 and E1 transport, 283–285
headend equipment, 254–255
home interface, 250–252
hybrid fiber-coax network, 231, 237, 256–260
modification of digital loop carrier for, 236–238
network interface device, identification, 258–259
network interface device, powering, 252–253
network interface device, transmitter level, 259
quality of service, 262–266

Cable telephony (*continued*)
redundant routing, 261–262
transmission methods, alternative, 262
Cable television
band plan, 393–402
changes in, 3–4
connection problems, 906–917
historical development of, 10–14
service reliability and availability, 790–792
Call blocking, 262
Call concentration, 263–266
Call processing, 234–235
tones, 242, 243
Canadian Research Council (CRC), 63
Captain, 60
Carrier null, 418
Carrier Sense Multiple Access with Collision Detection (CSMA/CD), 210
Carrier-to-composite-noise (CCN) ratio, 852–853
Carrier to distortion, 485–487
Carrier-to-noise-plus-interference (C/(N + I) ratio, 852–853
Carrier-to-noise ratio
bit error rate versus, 169–170
calculating, 606–607
coaxial distribution system design and, 483–485
performance and, 624–625, 628–629
satellites and, 315–317
Carrier-to-thermal-noise ratio (C/N), defined, 15
Cassegrain, William, 308
Cassegrain reflector, 308–309
Cathode, 31
Cathode ray tube (CRT), 31
CBS (Columbia Broadcasting System), 63, 65–66
Ceefax, 60
Cells, 112
Centralized architecture, 759–765
Central limit theorem, 681
Certificates
authentication, 203
authority, 878
digital, 878–881
revocation, 881
Channel(s)
allocation, list of, 985–990
assignments, in aeronautical band plans, 400–402
barker, 924
coding, 180
dense wavelength division multiplexing, 561
flatness, 582–585
forward application transport, 847, 852
forward data channels, 853
map/mapping, 117, 924
NTSC, 852
out-of-band, downstream, 853–855
out-of-band, upstream, 854–860
rejection, 949
reverse data channels, 853
sharing, 255–256
Characteristic impedance, 427, 430–432
Chat/conferencing, 899
Chirp, composite second order and with dispersion, 537–538
with imperfect channel flatness, 582–585
Chirp, defined, 784
Chroma noise, 49
Chromatic (wavelength) dispersion, 522–523
CH. ¾ RF interface, 861
Churning keys, 881–882
Circuit switch, 193
CLASS (custom local area signaling service), 239
Classifiers, 202, 204
Client-server system, 267
Clipping distortion, 550–555, 631
Clock recovery, 158
Closed captioning (CC), 59, 101
Coarse wavelength division multiplexing (CWDM), 560, 651
Coaxial cable
advantages of, 426
attenuation, function of characteristic impedance, 430–432
attenuation, function of frequency, 427–430
attenuation, function of temperature, 430
characteristic impedance, 427, 430–432
defined, 426
diameters, 437
drop, 438–440
early use of, 11
failure rates for, 805–806
hard line, 437
home networks, 972
precision of match, 433–436
selection of, 436–440
shielding effectiveness, 440
size limitations, 432–433
structural return loss, 433–434
time domain reflectometry, 434–436
trunk and feeder, 437–438
wavelength, 432
Coaxial distribution systems
bandwidth, 737
carrier to distortion, 485–487
carrier-to-noise ratio, 483–485
example, 736–739
legs, multiple, 740–741
noise-distortion trade-off, 487–490
powering, 737–739
service drop, 496–506
serving area size, 737
signal level management, 493–496
signal level stability, 496
system powering, 490–493

Coaxial systems, limitations of, 16–18
Coaxial technology
amplifiers, 440–469
cable, 425–440
directional couplers, 470–472
passive devices, 470–477
power supplies, 477–478
splitters, 472–474
taps, 474–477
Codecs, 273
Code division multiple access (CDMA), 174–175, 262
Coding
advanced video, 105–107
channel, 180
entropy, 93, 111
gain, 708
interleaving, 127–128, 180
linear predictive, 273, 330
of moving images/interframe, 93–100
of still images/intraframe, 79–93
transform, 84–90
Coefficient matrices, 95
Collision detection, 210
Color television, 47–56
compatibility, 49, 55–56
components, 82
human visual system response to, 47–48
I signal, 50–53
Q signal, 50–53
rectangular versus polar systems, 53–55
Comb generation technique, 353–354
Common application language (CAL), 504
Common path distortion (CPD), 676, 685
Communications Data Service, Inc., 944
Community antenna television (CATV), 11
Companded, signal, 57
Compatibility
analog television, 27–28, 47, 49
definitions, 930
description of problems, 925–930
digital cable, 131–132
high-definition television, 73
importance of, 918
legal issues, 930–934, 951–952
problems with, 918–925
standards, 934–937, 951–952
Competitive local exchange carriers (CLECs), 5
Complementary code keying (CCK), 976
Composite intermodulation noise (CIN), 455–456, 489, 585–587, 630–631
Composite reverse signal, 680–681
Composite second order (CSO), 443–448
caused by chirp with dispersion, 537–538
caused by chirp with imperfect channel flatness, 582–585
distortion, 588–589
performance and, 626–627
Composite triple beat (CTB), 448–454, 626–627
Compression
digital transmission and data, 329–330
in voice circuits, 273–274
Conditional access
MPEG-2 systems transport, 119
sharing, 893–894
Conference calling, 279–280
Congestion control, 226–227
Connection, residential gateway and, 960, 961
Connectionless service, 191–193
Connection-oriented service, 190–191, 193
Connection problems
for consumers, 906–917
set-top terminal problems, 915–916
splitting of cable signal, 912–914
TV to VCR to cable, 908–910
TV with built-in A/B switch, 914–915
two set-top terminal case, 910–912
Connectors
cleaning issues, 530–531
failure rates for, 804–805
ultra- versus angle-polished, 529–530
Constant bit rate (CBR), 389–391, 710
Constant power density, 691
Constellation diagram, 146–148
gain compression and, 162
interference and, 160–162
Consumer electronic products
features of, 905–906
frustration when making connections to products, 906–917
as an industry, 922–923
interface issues, 21–23, 918–925
terminal, 7–8
Consumer Electronics Association (CEA), 848, 868
Consumer Electronics Manufacturers Association (CEMA), 503, 936, 946
Content channel key, 882
Contention protocols, 189–190
Copy protection
algorithms, symmetric and asymmetric, 882–883
analog, 864
certificate revocation, 881
DES encryption algorithm, 884–888
Diffie-Hellman key exchange protocol, 883–884
digital signatures and certificates, 878–881
DTCP, 890–891
DVI, 888–890

Copy protection (*continued*)
encryption, 875–876
HDCP, 888–890
IEEE 1394, 890–891
keys, churning, 881–882
keys, public/private, 877–878
POD module, 891–892
Crimp ring, 500
Cross modulation (XMOD) distortion, 454–455, 486
Cross-phase modulation (XPM), 565
with fiber dispersion, 567–569
with transmission slope, 572–574
Cross-polarization modulation, 565
Crosstalk, 67, 145, 784–785
direct pickup, 65–66, 619
wavelength division multiplexing and, 565–582
Cryptography, 72
CSO. *See* Composite second order
CTB. *See* Composite triple beat
Custom calling service (CCS), 238
Custom local area signaling service (CLASS), 239
Custom messaging service (CMS), 238
Cutoff wavelength, 432
CWDM. *See* Coarse wavelength division multiplexing

Data encryption standard. *See* DES
Data link layer
DOCSIS, 199–200
MAC, 199
OSI, 186–187
Data Over Cable Systems Interface Specification (DOCSIS), 8
Data scrambling, 158
Data signal, spectrum of, 141–142
Data striping, 388
Data terminal equipment (DTE), 213–214
dBx Corp., 57
DCL. *See* Digital loop carrier
dc restoration, 343
DEC (Digital Equipment Corp.), 210
Decision boxes, 149–150
Decision threshold, 147
Decoding
audio, 109
interframe, 94
Delay equalizer filter, 339
Delays, 274–275
group, 165, 339
Demand-side management (DSM), 730
Demodulation
aural carrier, 360–362
depth of modulation (DOM), 415–417
envelope detection, 359, 363–368
intermediate frequency processing, 357–358
Nyquist slope filter, 362–363
RF input section, 357
synchronous detection, 359–360, 363–368
video output, 360
visual carrier, 359–360
Denial-of-service attack, 969
Denial technology
See also Copy protection
changes in, 821
interdiction, 825–827
negative traps, 822–823
non-set-top, 822–827
positive traps, 823–825
set-top program denial, 827–836
set-top terminal, 836–842
sync suppression scrambling, 827–836
Dense wavelength division multiplexing (DWDM), 531, 560
channels, 561
Depressed-clad, nondispersion-shifted fiber, 519
Depth of modulation (DOM), 413–417
Descrambler/descrambling
boxes, 20–21
defined, 821
dual, 905
synchronizing the, 830–831
waveform, 832–834
DES (data encryption standard) encryption algorithm, 884–888
Dialing
dual-tone multifrequency, 241–242
rotary, 241
Didon, 60
Dielectric constant, 510–511
Differential encoding, 158–159
Differential gain, 67
Differential phase, 67
Differentiator, 37
Diffie-Hellman key exchange 883–884
DigiCipher, 74
Digideck, 62, 63
Digital audio compression
Dolby AC-3, 110–111
MPEG-1, 111–112
MPEG-2, 112
principles of, 108–110
Digital broadcast television, responsibility in transition to, 6
Digital compression
applications, early, 71
audio, 108–112
broadcast digital television, 72–78
interest in, 71–72
lossy, 78
signal processing, 75–78
video, 78–108
Digital Content Protection, LLC, 888
Digital data, 137
Digital interfaces, 865–875
Digital loop carrier (DCL)
defined, 235–236
interfaces, 233, 236

modification of, for cable switched circuit telephony, 236–238
Digital modulation
bidirectional, 774–775
combining analog and, 772–773
difference between analog and, 768–769
receivers, 771–772
transmitters, 769–771
transmitters, low-frequency content removal in, 773–774
Digital program insertion (DPI), 122–124
Digital repeating ring architecture, 745–747
Digital signal processing (DSP), 172
Digital signals
advantages of, 72, 76–77
measuring, 175–180
processing, 75–78
Digital signatures and certificates, 878–881
Digital subscriber line (DSL), 4
Digital television, broadcast, 72–78
Digital-to-analog converters, 672
Digital transmission, 329–330
data compression and, 329–330
Digital transmission content protection. *See* DTCP
Digital versatile/video disk (DVD), 78
Digital video compression (DVC), 78–108
applications, early, 78–79
B frames and groups of pictures, 98
block formation, 84
coding of moving images, 93–100
coding of still images, 79–93
digital artifacts, 100–101
entropy coding, 93
field-based prediction, 98–100
fractal, 107
high-definition television, 102
motion compensation, 96–98
MPEG standards, 102–107
picture differencing, 93–95
principles of, 79
quantization, 90–91, 108
run length encoding, 91–92
sampling and preprocessing, 80–83
transform coding, 84–90
vector quantization, 108
vertical blanking interval, 101
wavelet, 107
Digital video interfaces
asynchronous serial, 376–378
Ethernet, 378
Digital Video Service Multiplex and Transport System for Digital Cable specification, 132
Digital video services
broadcast, 716
interactive, 716–717
private, 717
Digital Video Subcommittee (DVS), 848
Diplex filters
defined, 650
group delay and, 641–642, 652–653
midspan, 689
Dipole antennas, 288–293
Dipole patterns
E plane, 290–291
H plane, 291
Direct broadcast satellite (DBS) systems, 71
Direct feed
AM sublow band, 328–329
AM supertrunk, 328
baseband transmission, 327–328
digital transmission, 329–330
FM supertrunk, 329
Directional couplers, 470–472
Directivity, 470–472
Direct pickup (DPU) interference, 65–66, 619, 943–946
backfeed, 948
Direct sequence spread spectrum (DSSS), 175, 975–976
Direct-sound system, 57–58
Discrete cosine transform (DCT), 94, 330
8-by-8 blocks, 87–88, 89
Discrete interfering signal probability (DISP), 677
Discrete wavelet multitone (DWMT), 174
Dispersion
chromatic (wavelength), 522–523
composite second order caused by chirp interacting with, 537–538
material, 515, 522
modal, 517
polarization mode, 523
self-phase modulation interacting with, 525–528
waveguide, 522
wavelength, 514–516
Distortion
amplifier, 442–456
calculating, 607
clipping, 550–555, 631
coaxial distribution system design and, 485–487
common path, 676, 685
composite intermodulation noise, 455–456
cross modulation, 454–455, 486
even-order (composite second order), 443–448
floor, 442
noise-distortion trade-off, 487–490
odd-order (composite triple beat), 448–454
small-signal, 549–550
Distributed architecture, 756–758
Distributed feedback (DFB) laser diodes, 534–538, 669–673, 695–696

Distributed networks, 19–20
Distribution amplifiers, 463
Distribution system, 16
DIX standard, 210
DOCSIS
  data link layer, 199–200
  modems, 167, 183
  protocol, 194–207
  quality of service, 200–202, 203–207
  reference model, 194–196
  RF characteristics, 196–199
  security, 203
Dolby Digital audio, 893
  AC-3, 110–111
Dot crawl, 55–56
Double sideband modulation, 43
Downstream
  bandwidth, 702–703
  direction, 11
  features, 196, 197, 198
  frequency partitioning, 651–652
  out-of-band channels, 853–855
  signal level management, 493–494
Drop cable, 16, 438–440, 497
  See also Service drop
  shielding integrity, 499
Drop filters, interference mitigation and, 686–687
Drops, reflections in, 641
DS1 transport, 283–285
DTCP (digital transmission content protection), 890–891
DTV Applications Software Environment (DASE), 894
Dual conversion, 355–357
Dual descrambler, 905
Dual-tone multifrequency (DTMF) dialing, 241–242
DVI
  copy protection, 888–890
  interface, 865–868
DWDM. *See* Dense wavelength division multiplexing
Dynamic Host Configuration Protocol (DHCP), 200, 222, 963
Dynamic service, 202, 205–207

Earth station
  design, 326
  geometry, 311–312
  performance computations, 313–322
  receiving equipment, 373–376
Echoes, 165
  cancellation, 243, 275–276
  telephone, 242–244
E-commerce, 899
Edison, Thomas, 32
EIA. *See* Electronic Industries Association
8-phase shift keying modulation, 151–152
8-vestigial sideband modulation, 153–154
EIRP (effective isotropic radiated power), 313, 594
Electron gun, 29, 31
Electronic Industries Association (EIA), 368, 848
  captioning technical standards and, 59
  EIA-IS-23, 946
  EIA-IS-105, 937
  EIA-23, 852, 951
  EIA-542, 397–404, 630, 936
  EIA-563, 936
  EIA-679, 938
  EIA-708, 101
  home wiring system standards and, 503–504
Electronic program guides, 837, 898, 928
Email, 899
Emergency alert system, 124
Emergency calling, 282–283
EnCamera Sciences Corp., 63
Encoding
  A-law, 247, 273
  audio, 109
  differential encoding, 158–159
  G.711, 273
  interframe, 94
  Manchester, 140–141
  mu-law, 247, 273
  run length, 91–92
Encryption, 203
  See also Copy protection
  defined, 821
  description of, 875–888
Endpoints, 270, 271, 272
End-to-end fiber optics, link performance of, 548–555
Entitlement management messages (EMMs), 853
Entropy coding
  audio, 111
  video, 93
Envelope, 41, 184
  detection, 359, 363–368
E1 transport, 283–285
Equalizers, 494
Equipment reliability, 713
Erbium-doped fiber amplifiers (EDFAs), 540–542
Error amplifiers, 469
Errors
  bit error rate, 142, 169–170, 177–178
  correction, 689–690
  even parity, 77, 125
  forward error correction, 78, 125–128, 180
  identifying, 77–78
  parity detection, 77, 125
  RMS, 305–306
Error vector magnitude (EVM), 179
Ethernet
  addressing, 216
  development of, 209–210
  EPONs, 786
  frames, 215–216
  home networks, 971
  interconnecting devices, 213–215
  Internet Protocol embedded in, 218
  1000BaseT, 211

1000Base-X, 212
10/100BaseT, 210–211
twisted-pair wiring, 212–213
video transmission, 378
European Broadcast Union (EBU), 64
European Computer Manufacturers Association (ECMA), 900
European television
color, 67
flicker, 33
UHF frequencies, 68
Even-order distortion, 443–448
Even parity, 77, 125
Exhaustive search, 881
Extended data services (EDS), 59
Externally modulated transmitters continuous wave sources, 539–540
Extinction ratio, 769
Eye diagram, 159–160

Fabry, Charles, 533
Fabry-Perot laser diodes, 533–534, 668–669, 690–694
Failure rates, 713–714
See also Network analysis
defined and calculated, 792–795
single points, 715–716
Faraday rotation, 323–324
Far field signals, 289–290
Fast Fourier transform (FFT), 675
F connectors, 499–500
Federal Communications Commission (FCC)
cable antenna relay service (CARS) and, 18, 935–936
Cable Act (1992), 903, 923, 925, 930–933
cable ready, 951–952
captioning technical standards and, 59
closed captioning and, 101
consumer terminal equipment and, 7–8
digital data transmission and, 62
early standards for televisions, 27–28
high-definition television and, 73
methodology for predicting signal strength, 302–303
performance standards and, 619–620, 621
signal leakage standards and, 9
set-top terminals and, 847
Feed-forward amplifiers, 469
Feed-through F connectors, 500
Fiber-deep architecture
description of, 775–780
distance limitation, 780–785
standards, 786
Fiber nodes
multiple coaxial distribution legs, 740–741
redundant multiple, 739–740
redundant signaling, 740
size and architecture, 752–754
split-band multiple, 739
Fiber optic links, 741
performance, end-to-end, 548–555
ring, analog shared, 744–745
ring, digital repeating, 745–747
sheath ring, 743–744
star, 742–743
Fiber optics
amplifiers, 540–542
clipping distortion, 550–555
early use of, 13
end-to-end performance, 548–555
failure rates for, 805
graded index, 517
interactions among transmitters, fibers, and receivers, 544–547
introduction of, 18–19
light absorption, 514
management, 411–413
multimode, 516–518
passive networks, 528–531
Rayleigh scattering, 514, 545–547
receivers, 542–544
reflection and refraction, 512–514
step index, 517
velocity of propagation and wavelength, 510–512
wavelength dispersion, 514–516
Fiber optics, linear distributed feedback laser diodes, 534–538
externally modulated transmitters, continuous wave sources, 539–540
Fabry-Perot laser diodes, 533–534
general characteristics, 532–533
Fiber optics, single-mode
applications, 518
attenuation, 519–521
attenuation in bent fibers, 523–524
chromatic (wavelength) dispersion, 522–523
polarization mode dispersion, 523
self-phase modulation interacting with dispersion, 525–528
stimulated Brillouin scattering, 524–525
structure of, 518–519
Fiber-to-the-curb (FTTC) architecture, 524, 735, 752, 775–776, 778
Fiber-to-the-home (FTTH) system, 776–780
Field-base prediction, 98–100
Field factor, 606
Filtering
frequency shift keying and, 140
methodology, 155
peak-to-average ratio and, 164–165

Filters
adaptive, 198
matched, 155
raised cosine, 155
root raised cosine, 155
sine squared, 155
Fink, Donald, 39
Firewall, 968, 969–970
FireWire™, 868
Flat loss, 472
Flicker fusion frequency, 33
Flickers, 32–33
FM band plans, 402
FM receiver characteristics, 317–319
FM supertrunk, 329
FM terrestrial microwave, 327
Footage, fiber versus cable, 744
Footpath, satellite, 313
Forward application transport (FAT) channel, 847, 852
Forward data channels (FDCs), 853
Forward error correction (FEC), 78, 125–128, 180
Forward scattering, 511
Fourier series equations, 38, 85–87
Four-wave mixing (FWM), 565, 576–577
Fractals, 107
Fragmentation, 202
Frames
audio, 110
cable telephony, 256
Ethernet, 215–216
I, 95
P, 95
Free space loss, 606
Free space propagation, 300–301
Frequency
agility, 337
assignment, 256, 651–652
coefficients, 88
deviation, 138
response variation, 461–462, 501, 642
Frequency diverse QAM (FDQAM), 971
Frequency division multiplexing (FDM), 14, 40, 125, 172
Frequency domain view, laser clipping and, 681–683
Frequency hopping spread spectrum (FHSS), 175, 689, 974–975
Frequency modulation, 45–46
Frequency shift keying (FSK), 138–141
Fresnel zone, 599
calculating, 600
Fretting corrosion, 685
Fringe reception, 941
Full-Service Access Network (FSAN), 786
Fusion splicing, 529

Gain compression, constellation diagram and, 162
Gain-flattened amplifiers, 563–564
Gaussian white noise, 28
General Instrument, 74
Ghost reduction, antenna, 296–300
Gibbs phenomenon, 101
G-line, 426
Graded index fibers, 517
Gray, ?FIRST NAME, 39
Group delay, 165, 339
amplifiers and, 456–457
due to diplex filters, 641–642, 652–653
performance and, 632–642
service drop and, 501
upstream, 652–653
Groups of pictures (GOP), 98, 119
G.711 encoding, 273

Half duplex communications, 266
Hall's equation, 301
Hanover bars, 67
Hard-line cable, 437
Harmonically related carrier (HRC) technique, 351
band plan, 400
comb generation, 353
Harmony agreement, 893–894
Haystock response, 363
HDCP (high-bandwidth digital copy protection), 888–890
Headends, 11, 14
See also Antennas; Satellites
advertising insertions, 385–387
cable telephony equipment, 254–255
demodulation, 357–368
digital video interfaces, 376–378
failure rates for, 806, 809
fiber-optics management, 411–413
FM terrestrial microwave, 327
handling signals, 378–385
modulation, 341–350
phase locking of carriers, 350–357
processors for, 333–341
quality tests, 413–422
RF management, 402–410
in satellite earth station receiving equipment, 373–376
switched video, 747–750
in television stereo systems, 368–373
video on demand, 388–391
Header, 184
Internet Protocol, 218–220
High-definition television (HDTV)
block formation, 102
compatibility, 73
digital compression and, 71, 73–74
digital video compression and, 102
early developments, 73
sampling, 102
signals, 73–74

High-pass filter, 37
High-split networks, 703
Hitachi, 890
Home Box Office (HBO), 11
Home networks
  applications, 956–957
  CableHome, 979–982
  configuration, 960–961, 966–968
  connectivity, 960, 961
  installation of, 958
  multiplexing and address management, 960, 961–966
  performance management, 960–961, 966–968
  quality of service, 961, 970
  requirements for residential gateway, 959–968
  security, 961, 958–959, 968–970
  support and maintenance, 958
  wireless networks, 972–979
  wired networks, 970–972
HomePlug Powerline Alliance, 971–972
HomePNA, 971
Home wiring systems, integrated, 503–506
Horizontal blanking interval (HBI), 32
Horn antennas, 311
Host
  cable input, 852–853
  description and block diagram of, 849–852
  use of term, 849
Host digital terminals (HDTs), 250, 755
Host-POD Interface specification, 118–119, 132
HRC. *See* Harmonically related carrier
H.323 protocol, 282
Huffman codes, 93
Hum
  audio, 381
  buildup, 493
  modulation, 459–461, 631
  video, 383–385
Hybrid fiber-coax (HFC) network, 71, 231, 237
  frequency assignment, 256
  marshaling and provisioning, 256
  network analysis of, 810–817
  network interface device, identification, 258–259
  network interface device, powering, 252–253
  network interface device, transmitter level, 259
  performance example, 627–645
  serving area size, 737
  TDM/TDMA transmission, 257–258
  time slot assignment, 258
  timing offset, 259–260

IBC, 368
IBM, 210, 884
Iconoscope, 31
IEEE
  definition of noise, 483
  802.3, 210
  802.3ah, 786
  802.11, 974–976
  802.11a, 978–979
  802.11b, 976–977
  802.11g, 977–978
  1394 copy protection, 890–891
  1394a interface, 868–873
IF. *See* Intermediate frequency
I frame, 95
i.LINK™, 868
Image rejection, 949–950
Image response, 393–395
Imperfect demultiplexer isolation, 566–567
Impulse-pay-per-view (IPPV) programming, 649, 838–839
In-band addressability, 838
Incidental wavelength modulation, 536–538
Incrementally related carrier (IRC) technique, 351
  band plan, 399–400
  comb generation, 353–354
Independent Broadcast Authority (IBA), 60
Index of refraction, 511
Information capacity, 702–712, 722–723
In-home equipment, 734, 849–853
In-phase color signal, 53
Instantaneous circuit data rate, 711
Integrated services digital network. *See* ISDN
Integrator, 35
Intel, 210, 890
Interactive digital video services, 716–717
Intercarrier frequency, 361, 414
Intercarrier-sound method, 57
Interdiction, 825–827
Interfaces, 233, 236, 249
  See also Copy protection; Digital video interfaces
  analog, 861–865
  cable services and, 903–905
  consumer electronics and, 905–906, 918–925
  digital, 865–875
  home, 250–252
Interference
  See also Noise
  antennas and, 296–300
  direct pickup, 65–66, 619, 943–946
  levels of, 942–943
Interferometric intensity noise (IIN), 545–547
Interframe coding (moving images), 93–100
  decoder, 94
  encoder 94
Interlaced scanning, 33–34
Interleaving, 127–128, 180
InterMedia Partners, 951
Intermediate frequency (IF)
  demodulation, 357–358
  processing, 344–345, 357–358
  in signal processors, 335–336

International Consultative Committee for Radio (CCIR), 64
International Standards Organization (ISO), 8, 102
International Telecommunications Union (ITU), 185, 304, 878, 893
Internet access requirements, 724–727
Internet Control Message Protocol (ICMP), 221
Internet Engineering Task Force (IETF)
  Real-Time Control Protocol, 276, 277
  Real Time Protocol, 276–277
Internet Group Management Protocol (IGMP), 221–222
Internet Protocol (IP), 216–221
  addresses, 221
  embedded in Ethernet, 218
  header, 218–220
  IPv4 versus IPv6, 218
Internet protocol (IP) telephony, 899
  codecs, 273
  compression in voice circuits, 273–274
  conference calling, 279–280
  delay, 274–275
  development of, 266–267
  echo cancellation, 275–276
  MGCP and TGCP standards, 272–274
  Real Time Protocol, 276–277
  User Datagram Protocol, 277
  voice activity detection, 274
  VoIP, description of, 267–271
  VoIP, makeup of packets, 277–279
  VoIP control protocols, 281–282
  VoIP signal, bandwidth requirements, 280–281
Internet service providers (ISPs)
  bundling of, 9
  reliability, 726–727
Intersymbol interference (ISI), 160, 165
  constellation diagram and, 160–162
Intraframe coding (still images), 79–93
Inverse discrete cosine transform (IDCT), 88, 89, 90, 94
Inverse Network Technology, 726
IRC. *See* Incrementally related carrier
IRD two-carrier stereo system, 368, 372–373
ISDN (integrated services digital network), 239–240
Isolation
  A/B switch, 948
  coupler, 471
  imperfect demultiplexer, 566–567

Jones, Bronwen, 15
Carl T. Jones Corp., 944, 945, 946

Kell, R. D., 45
Kell factor, 45
Key exchange, 203
Kinescope, 31
Kodak Corp., 64

Laser clipping
  frequency domain view, 681–683
  time domain view, 683–684
Lasers, upstream
  characteristics of, 668–674
  combining paths, 673–674
  data- or video-grade classification, 662–663
  discrete carrier testing and classification, 663–664
  noise power ratio testing and classification, 664–667
  specifying, 662–668
LATA (local access and transport area), intra and inter, 233
Latency, 180
L-C filters, 339, 349
Light absorption, 514
Light insertion and detection (LID) splicers, 524
Linear distortions, 165
Linear feedback shift register (LFSR), 883
Linear fiber optics. *See* Fiber optics, linear
Linear predictive coding (LPC), 273, 330
Line balance, 244
Line concentration, 728
Line controllers, 233
Line extenders, 463
Links. *See* Fiber optic links
Local exchanges, 232–233
Local origination, 330
Local oscillators, 337
Logical link control (LLC), DOCSIS, 199–200
Logical link layer, OSI, 187
Log periodic antennas, 295–296
Loop out, 344
Loss, defined, 720
Lossy compression, 78
Low-frequency enhancement (LFE), 110
Low-pass filters (LPFs)
  frequency shift keying and, 139
Low-split networks, 703

Mach-Zehnder devices, 539, 554
Macrobends, 524
Manchester encoding, 140–141
Mantissa, 111
Marshaling and provisioning, 256–260
Masking, 108
Matched-clad, nondispersion-shifted fiber, 519
Matched filters, 155
Material dispersion, 515, 522
Matsushita, 890
Maxwell's equations, 289
MCNS Holdings, 194
Mean time between failures (MTBF), 714, 794

Mean time to restore (MTTR) service, 714, 794–795
estimating, 807–808
Mechanical splices, 529
Media access control (MAC), OSI, 187
services provided, 199
time slots, contention for, 189
time slots, reservation of, 188
MEGACO protocol, 282
Mertz, ?FIRST NAME, 39, 49
Metcalf, Bob, 209, 210
Microbends, 523–524
Microreflections, 501–502, 631–632
Microwave, FM terrestrial, 327
Microwave signals
band frequencies, 595
channels, 596
link availability factors, 607–615
operational principals, 594–597
path design, 597–603
performance, calculating, 604–607
rain, effects of, 609–614
U.S. transmission regulations, 593–594
Middleware, 894–900
Midsplit networks, 703
Modal dispersion, 517
Mode partition noise (MPN), 534, 783–784
Modulation
See also Quadrature amplitude modulation (QAM),
amplitude, 40–43
analog television, 40–46
aural, 345–348, 417–421
automatic gain circuit, 342
baseband spectrum of a data signal, 141–142
biphase shift keying, 142–145
bit error rate, 142, 169–170
cross-phase, 565, 567–569, 572–574
cross-polarization, 565
dc restoration circuit, 343
defined, 40
depth of, 413–417
diagram of, 341
digital, 350, 767–787
digital transmission and, 125
double sideband, 43
8-PSK, 151–152
8-VSB, 153–154
frequency, 45–46
frequency shift keying, 138–141
hum, 459–461, 631
incidental wavelength, 536–538
optical, index, 565
peak white clipper, 343
process, 343–344
quadrature, 50
quadrature phase shift keying, 145–148
RF transmission, 154–168
self-phase, 525–528
signals, measuring, 175–180
single sideband, 43
sound subcarrier trap, 342–343
spectral efficiency, 138
spectrum sharing, 170–175
vestigial sideband, 43, 348–350
video switch, 342
Modulation error ratio (MER), 178–180
measuring, 178–180
Moore, Gordon, 76
Mosquito noise, 101
Motion compensation, 96–98
Motion vector, 96
Moving pictures, 32–33
MPEG (Motion Picture Experts Group) standards
video compression, 102–103
MPEG-1, 103–104
MPEG-1 audio, 111–112
MPEG-2, 72, 74, 893
audio, 112
conditional access, 119
description of, 104–105
multiplexing, 114–116
navigation, 116–119
splicing, 119–120
systems layer, 112–120
timing distribution, 119
transport packet format, 115
video compression, 75, 78
MPEG-4, 105
Part 10, Advanced Video Coding (AVC), 105–107
mu-law encoding, 247, 273
Multilayer architecture, 754–756
Multimedia Cable Network System, 194
Multimedia Gateway Control Protocol (MGCP), 269–274
Multimedia Home Platform (MHP), 894
Multimode fiber optics, 516–518
Multipath, 607–609
Multiple feeds, satellite, 310–311
Multiple-program transport stream (MPTS), 114, 116
Multiple sub-Nyquist sampling encoding (MUSE), 73, 74
Multiplexing
frequency division, 14, 40, 125, 172
MPEG-2 systems transport, 114–116
residential gateway and, 960, 961–963
space division, 704
statistical, 120–122, 390, 710–711
time division, 112, 125, 172, 247, 257–258, 710
wave division, 329, 531, 559–590
MUSICAM, 111

Narrowband transmission channel, 729
Narrowcast video services requirements, 721–723
NASA, 63
National Association of Broadcasters (NAB), 65

National Bureau of Standards (NBS), 58–59
National Cable Television Association (NCTA), 906, 936, 946
National Data Broadcasting Committee (NDBC), 62
National Institute of Standards and Technology (NIST), 59, 884
National Renewable Security Standard (NRSS), 937–938
National Television Systems Committee (NTSC)
  active picture region, 83
  color system, 67
  standard created by, 11, 27, 28, 852
Navigation, MPEG-2 systems transport, 116–119
NBC, 59–60
NCI, 59
Near field signals, 289–290
Near instantaneous amplitude modulation (NICAM) stereo system, 368, 371–372
Near video on demand (NVOD), 13, 747–750, 839
Negative traps, 20, 822–823
Net gain, 494
Network
  access, by unaffiliated third parties, 5–6
  availability, 714
  call signaling (NCS), 270
  failure rate, 713–714
  high-split, 703
  low-split, 703
  midsplit, 703
  stream, 123
Network address port translation (NAPT), 965–966, 969–970
Network address translation (NAT), 963–965
Network analysis
  assigning components to subdivisions, 800
  determining component failure rates, 800–807
  estimating mean time to restore, 807–808
  headends and hubs, 809
  logical subdivision of, 799–800
  two-way systems and, 809–810
Network interface devices (NIDs), 237–238, 718
  failure rates for, 806
  identification, 258–259
  powering using backup battery, 253
  powering using center conductor, 252–253
  powering using twisted pair, 253
  transmitter level, 259
Network interface unit (NIU), 718
Network layer, OSi, 187
Networks, passive, 528
  attenuators, 531
  connectors, 529–531
  signal splitters, 531
  wavelength division multiplexers, 531
A.C. Nielsen Co., 62, 63
911 emergency calling, 282–283
Nipkow, Paul, 29
Nodes
  fiber, 739–741
  scalable, 651
  two-way, 650–651
  use of term, 650
Noise
  See also Carrier-to-noise ratio; Chirp; Crosstalk
  amplifier, 441–442
  bandwidth, 176
  blocking, 100
  carrier-to-composite-, 852–853
  carrier-to-noise-plus-interference, 852–853
  chroma, 49
  composite intermodulation, 455–456, 489, 585–587, 630–631
  defined, 483–484
  -distortion trade-off, 487–490
  end-to-end fiber optics and, 548–555
  figure, 441
  floor, calculating, 440–441
  Gaussian white, 28
  interferometric intensity, 545–547
  mitigation techniques, 685–690
  mode partition, 534, 783–784
  mosquito, 101
  phase, 547
  quantization, 81, 246
  relative intensity, 534–536
  shot, 780–781
  signal-to-noise ratio, 170, 247, 317, 782–783
  telephone, 242–244
  thermal, 14–15, 484, 781
  upstream, 674–679, 685–690
  white, 585–587
Noise power ratio (NPR), upstream lasers and, 664–668
Nonpredicted, 95
Non-real-time polling service (nrtPS), 202
North America Presentation Layer Protocol System (NAPLPS), 60–61
NTSC. *See* National Television Systems Committee
Nyquist, Harry, 80
  bandwidth, 245
  slope filter, 362–363
  theorem, 80, 87

OAM&P, 255
Odd-order distortion, 448–454
Off-air bands, 398–399
Off-air processing, 339–341
Off-air reception, 287
  antennas for, 288–300
  propagation of signals, 300–303
Offset feeds, 309–310

On-demand/interactive video services requirements, 723–724
OpenCable, 130–131, 848, 937
OpenCable Applications Platform (OCAP), 848
  1 specification, 894, 895–900
  2 specification, 900
Open loop, 496
Open Systems Interconnection (OSI) protocol
  application layer, 188
  data link layer, 186–187
  network layer, 187
  physical layer, 185–186
  presentation layer, 187–188
  session layer, 187
  transport layer, 187
Open video system (OVS) operator, 3
Optical distribution architecture, 759–765
Optical Kerr effect (OKE), 565
  with polarization dependent loss, 574–576
Optical line termination (OLT), 776
Optical link information capacity, 702
Optical modulation index (OMI), 565, 585–586
Optical network termination (ONT), 776
Oracle, 60
OR operation, 158
Orthogonal frequency division multiple access (OFDMA), 173
Orthogonal frequency division multiplexing (OFDM), 165, 173–174, 978
Oscilloscope
  audio monitoring using, 419
  depth of modulation using, 415–417
Outage
  See also Service reliability and availability
  defined, 720
  Outage Reduction, 719–720, 790–792
Out-of-band addressability, 837
Out-of-band (OOB) channels
  downstream, 853–855
  upstream, 854–860
Outside cable entrance facility (OCEF), 411, 413
Overbuilder, defined, 3
Overhead, 78

PacketCable, 272, 273, 276, 848–849
Packet data services requirements, 724–727
Packet identifier (PID), 114, 116–118
Packetized elementary stream (PES) packet, 113–114, 115
Packets, 112, 256
  VoIP, 277–279
Pads, 494
Pairing, 35
Panasonic, 890
Parabola antennas, 305–307
Parallel hybrid amplifiers, 469
Parity detection, 77, 125
Passive optical networks (PONs), 776–777
Path loss, 313–314
Paths, microwave
  azimuth, 599
  clearances, 599–603
  end coordinates, 597–598
  length, 598–599
  multi, 607–609
Payload header suppression (PHS), 200–202, 281
Pay-per-view (PPV) programming, 721, 838, 898–899
Peak-to-average ratio, 162–165
Peak white clip, 343
Peering system, 267
Performance
  architectures and, 701–718
  calculating microwave, 604–607
  composite second order, 626–627
  composite triple beat, 626–627
  carrier-to-noise ratio, 624–625
  hybrid fiber-coax network example, 627–645
  primary impediments, 622–623
  quality standards and requirements, 617–622
  of receivers, 642–643, 946–951
Peripheral component interface (PCI), 195
Perot, Alfred, 533
Persistence of vision, 29
Personal video recorder (PVR), 389, 894
P frame, 95
Phase alternating at line (PAL), 67–68, 83
Phase locking
  circuits, diagram of circuit, 352
  comb generation technique, 353–354
  defined, 351
  dual conversion, 355–357
  harmonically related carrier technique, 351, 353
  incrementally related carrier technique, 351, 353–354
  for processors or modulators, 351–357
  reasons for, 350–351
Phase noise, 547
Phase shift keying modulation 8-, 151–152
  quadrature, 145–148
Phasors, 40
Photons, 510
Physical layer, OSI, 185–186
Picture carrier levels, measuring, 421–422
Picture differencing, 93–95
Picture elements, 28
Picture-in-picture (PIP), 906, 926–927
Picture quality, 128–129

Pigtails, 411
Pirate, use of term, 821–822
Pixels, 28
Plant unavailability analysis, 676–679
Plug-and-play device, 966
POD (point-of-deployment)
  Copy Protections System specification, 132
  security module, 847, 891–892
POD-host interface. *See* Host-POD Interface specification
Point-of-entry (POE) equipment, 734–735
Point-of-entry signal processing, 717–718
Polarization
  dependent loss, 574–576
  satellite, 304, 322–324
  types of, 289
Polarization mode dispersion (PMD), 523
Ports, 223
Positive traps, 20, 823–825
Postheader, 184
POTS (plain old telephone service), 238
Power pack, use of term, 459
Power supply (powering)
  coaxial distribution system design and, 490–493, 737–739
  failure rates for, 803–804
  home networks, 970–979
  network interface device, 252–253
  switching transient effects, 492–493
  types of, 477–478
  use of term, 459
Preemphasis, 345–347
Presentation layer, OSI, 187–188
Primary-line service, 4
Prime focus feed, 307–308
Private keys, 877–878
Program and System Information Protocol (PSIP), 117–118, 340
Program clock reference (PCR), 119
Program insertion, 122–124
Programming by competing network operators, access to, 8
Propagation
  antenna height, effects of, 301–302
  free space, 300–301
  of off-air signals, 300–303
  wavelength and velocity of, 510–512
Protocols
  Asynchronous Transfer Mode, 208–209
  concept of, 183–185
  connectionless service, 191–193
  connection-oriented service, 190–191, 193
  contention/aloha, 189–190
  defined, 183
  DOCSIS, 194–207
  Ethernet, 209–216
  H.323, 282
  Internet, 216–222
  Layer 4 Transport, 222–227
  MEGACO, 282
  Multimedia Gateway Control, 269–274
  Open Systems Interconnection, 185–188
  Program and System Information, 117–118, 340
  Real Time, 276–277
  Real-Time Control, 276, 277
  Session Initiation, 282
  Simple Network Management, 200, 966, 967–968
  Transmission Control, 224–226
  Trunking Gateway Control, 269–274
  User Datagram, 224, 269, 277
Provisioning, 256–260
Psychoacoustic masking threshold, 108, 111
Public, educational, and governmental (PEG) access issues, 10
Public Broadcasting System (PBS), National Datacast Network, 61
Public keys, 877–878
Public rights-of-way, access to, 10
Public switched telephone network (PSTN), 190, 266
Pulse distortion, 165
Push-pull amplifiers, 469

Quadrature amplitude modulation (QAM), 21, 50
  bandwidth, required, 156
  binary numbering, 148
  16-QAM, 148–150
  32-QAM, 151
  64-QAM, 130, 150–151
  256-QAM, 130, 150–151
Quadrature color signal, 53
Quadrature detection, 50
Quadrature phase shift keying (QPSK), 145–148, 321
Quality
  picture, 128–129
  service drop and, 498–499
  signal, 63–66, 128, 714–715
Quality of Service (QoS), 970
  CableHome, 981
  cable telephony and, 262–266
  DOCSIS, 200–202
Quality tests
  aural modulation, 417–421
  depth of modulation (DOM), 413–417
  picture/sound carrier levels, measuring, 421–422
Quanta, 510
Quantization
  audio, 111
  noise, 81, 246
  vector, 108
  video, 90–91, 108
Quantizer matrix, 91

Rabbit ears, 299, 832–834
Rain, microwaves affected by, 609–614
Raised cosine filters, 155
Raster, 31
Rayleigh scattering, 514, 545–547
Real-time polling services (rtPS), 202
Real-Time Control Protocol (RTCP), 276, 277
Real Time Protocol (RTP), 276–277
Receivers
  broadcast versus broadband, 941–942
  digital and analog, 771–772
  direct pickup interference, 65–66, 619, 943–946
  failure rates for, 804
  fiber optical, 542–544
  interference from internal wiring of, 643–644
  interfering signals from, 642–643
  performance of, 642–643, 946–951
  use of term, 650
Redundant array of independent/inexpensive disks (RAID), 388
Redundant multiple fiber nodes, 739–740
Redundant networks, effects on service reliability and availability, 795–796
Redundant options, optical, 761–763
Redundant routing, 261–262
Redundant signaling, 740
Reencoding, 122
Reference carrier, 351
Reflections, 165
  in drops, 641
  due to mismatches in trunk and distribution plant, 634–641
  fiber optics, refraction, and, 512–514
Refraction
  fiber optics, reflection, and, 512–514
  index of, 511
Regulatory issues
  access to networks by unaffiliated third parties, 5–6
  access to utility poles and public rights-of-way, 10
  digital signal quality, 9
  market power issues, 7–9
  must carry/retransmission consent issues, 910
  public, educational, and governmental access issues, 10
  responsibility for transition to digital broadcast television, 6
  signal leakage standards, 9
Relative intensity noise (RIN), 534–536
Reliability. *See* Service reliability and availability
Remote controls, 925–926
  universal, 927–928
Repeaters, 76
Reservation, 188
Residential gateway (RG)
  defined, 959–960
  requirements, 960–968
Retrace, 32
Return. *See* Upstream
Reverse data channels (RDCs), 853
RF characteristics, 196–199
RF management, 402
  downstream, 403–407
  upstream, 407–410
Ring architecture
  analog shared, 744–745
  digital repeating, 745–747
Ringing, telephone, 240–241
RMS error, 305–306
Root raised cosine filters, 155
Routing, redundant, 261–262
RSA algorithm (Rivest, Shamir, and Adleman), 878–879, 882
Run length encoding, 91–92
Running disparity, 774

Sampling
  audio, 110
  video, 80–83, 102
  voice signal, 245
Satellite reception
  antennas for, 304–311
  azimuth, 311–312
  bands for, 304, 322–323
  carrier-to-noise ratio, 315–316
  Cassegrain reflector, 308–309
  digital, 321–322
  earth station design, 326
  earth station geometry, 311–312
  earth station performance computations, 313–322
  earth station receiving equipment, 373–376
  EIRP, 313
  elevation, 311–312
  frequency plans, 322–323
  horn antennas, 311
  multiple feeds, 310–311
  noise bandwidth, 315–316
  offset feeds, 309–310
  parabola, 305–307
  path loss, 313–314
  polarization rotation, 323–324
  prime focus feed, 307–308
  receiver characteristics, 317–319
  receiving stations, 324–326
  scrambling of signals, 320–321
  shrouds, 311
  slant range, 311–312
  threshold extension, 319
  ultra directive antennas, 311
Satellites
  footprint, 313
  G/T, 314–315
  life of, 303
  orbital slots, assignment of, 304

Satellites (*continued*)
pointing of, 311–312
polarization, 304
signals, early use of, 11
spacing between, 324–326
transponders, 304
Saturation, 53
Scalability, 730–731, 763–765
Scalable nodes, 651
Scaling constant, 296
Scanning
analog television, 28–34
interlaced, 33–34
photograph on rotating drum example, 28–29
spectrum of images, 37–39
Scattering
forward, 511
Rayleigh, 514, 545–547
stimulated Brillouin, 524–525, 780
stimulated Raman, 565, 569–572
Scientific notation, 111
Scrambling/scrambled
See also Copy protection; Denial technology
audio, 836
cable, 158
data, 158
defined, 821
of satellite signals, 320–321
sync suppression, 827–836
video inversion, 834–836
waveform, 831
SCTE. *See* Society of Cable Telecommunications Engineers
SECAM system, 68
Secondary-line service, 4
Second audio program (SAP) channels, 57
Security, 20–21
CableHome, 981–982
DOCSIS, 201, 203
firewall, 968, 969–970
home networks, 958–959, 968–970
Self-phase modulation (SPM)
interacting with dispersion, 525–528
Service bundling, 9
Service drop, 496
adding amplification, trade off of, 502–503
F connectors, 499–500
integrated in-home wiring systems, 503–506
microreflections, 501–502
quality issues, 498–499
reference structure, 497–498
signal leakage and ingress, 502
simple passive architectures, 497–502
standard, 498
transmission loss, 501
traps, 500–501
Service flows, 200, 202, 203–207
Service Information Delivered Out-of-Band for Digital Cable Television specification, 132
Service providers, 919–922
Service reliability and availability, 714–715
absolute versus user-perceived, 796–799
cable television, 790–792
definitions and calculations, 792–795
equipment, 713
failure rates, 803–810
hybrid fiber-coax network analysis for, 810–817
network, 714, 789–817
network analysis for, 799–810
redundancy, effects of, 795–796
telephony, 789–790
Session Initiation Protocol (SIP), 282
Session layer, OSI, 187
Set-top terminals (STTs)
authorization procedures, 838–839
classification, 837
connection problems, 906–917
denial technology for, 827–836
failure rates for, 806
functionality limitations due to, 23
in-band addressability, 838
out-of-band addressability, 837
signal path, 839–841
standards, 847–849
terminal control, 842
thin client, 901
TV display issues, 842
Shadow mask, 47
Shared terminal equipment, 735–736
Sheath ring architecture, 743–744
Shielding integrity, 22
Shot noise, 780–781
Siamese cable, 253
Sidebands, 42
defined, 344
double, 43
single, 43
vestigial, 43
Sidetone, 242
Signal(s)
See also Digital signals; Microwave signals; Satellite reception; Satellites
ancillary, 58–63
apportionment, examples of, 690–696
bandwidth, required, 156–157
composite reverse, 680–681
data, spectrum of, 141–142
far field, 289–290
leakage and ingress, 502, 946–947
leakage standards, 9
methodology for predicting strength of, 302–303
near field, 289–290
propagation of off-air, 300–303
quality, 63–66, 128, 714–715
quality tests, 413–422
rejection of, in upstream band, 950–951

reradiation of, 946–947
splitters, 531
voice signal sampling, 245
VoIP bandwidth requirements, 280–281
Signaling system number 7 (SS7), 233
Signal levels
automatic level control, 658–660
digitally modulated, 168
management, 493–496
measuring, 176, 655–658
stability, 496, 644–645
thermal gain control, 660
trap, 660–662
upstream, 493–496, 655–662
Signal loss. *See* Attenuation
Signal processing
point-of-entry versus user-terminal, 717–718
redundant, 740
single point versus distributed, 716–717
Signal processors
block diagram of, 334
digital off-air processing, 339–341
input section, 334–335
intermediate frequency, 335–336
local oscillators, 337
off-channel rejection, 338–339
output section, 336–337
pass band, 338–339
Signal reception
direct feed, 327–330
FM terrestrial microwave, 327
local origination, 330
off-air, 287–303
satellite, 303–326
Signal-to-noise ratio, 170, 247, 317, 782–783
Signatures, digital, 878–881
Simple Network Management Protocol (SNMP), 200, 966, 967–968
Sine squared filters, 155
Single-mode fiber optics. *See* Fiber optics, single-mode
Single-program transport stream (SPTS), 116
Single sideband (SSB) modulation, 43
Skin depth, 428
Skin effect, 428
Skirts, 176
Snell's law, 512
Society of Cable Telecommunications Engineers (SCTE), 111, 112, 440, 936
DOCSIS protocol, 194–207
emergency alert system, 124
standards for channels (SCTE 55–1 and SCTE 55–2), 854–860
standards for drop components, 498–499
standards of, 129–130, 132, 847, 848
Society of Motion Picture and Television Engineers (SMPTE), 64, 120
Sockets, 223
Softswitch, 268–269, 271
SONET (synchronous optical network)
description of, 249–250
TR-303 interface, 236, 249
Sony, 890
Sound carrier levels, measuring, 421–422
Sound subcarrier trap, 342–343
Space division multiplexing (SDM), 704
S/P DIF (Sony/Philips digital interface), 873–875
Spectrum analyzer, 415, 421–422
Spectral efficiency, 138
Spectrum sharing
code division multiple access, 174–175
frequency division multiplexing, 14, 40, 125, 172
orthogonal frequency division multiplexing, 165, 173–174
time division multiple access, 171–172
time division multiplexing, 112, 125, 172
Splicing, MPEG-2 systems transport, 119–120
Split-band multiple fiber nodes, 739
Split-sound method, 57–58
Splitters, 472–474, 531
Spoofing, 827
Spreading code/sequence, 174
Spread spectrum, 46, 174–175
Standards, 129–132, 847–849, 934–937
Star architecture, 742–743
StarSight Electronic Program Guide (EPG), 61
Static service flows, 202
Statistical multiplexing, 120–122, 390, 710–711
Status monitoring, 250
Step index fibers, 517
Stereo sound, analog television, 57–58
Stern Telecommunications Corp., 944
Stimulated Brillouin scattering (SBS), 524–525, 780
Stimulated Raman scattering (SRS), 565, 569–572
Streaming media, 72, 78–79
Structural return loss (SRL), 433–434
Subscriber line audio-processing circuit (SLAC), 270
Subscriber/consumer relationship, 918–919
Subscriber line interface circuit (SLIC), 270
Subscriber loop, 240–244
Subscription programming, 838
Subscription video-on-demand (SVOD), 839
Surface acoustic wave (SAW) filter, 339, 349, 350

S-VHS interface, 862–863
Switch architecture, telephone digital 233–235
Switched broadcast, 710
Switched video headend structures, 747–750
Switched virtual circuit, 710
Switching matrix, 233
Symbol error rate (SER), 678
Synchronization (sync)
  analog television, 34–37
  detectors in receivers, 55
  separator circuits, 35
  suppression scrambling, 827–836
Synchronous code division multiple access (S-CDMA), 174, 198
Synchronous detection, 359–360, 363–368
Synchronous digital hierarchy (SDH), 249
Synchronous discrete multitone (SDMT), 174
Synchronous optical network. *See* SONET
System information, use of term, 853
System renewability messages (SRMs), 891

Taps, 474–477
  losses in, 494
  signal levels at, 660–662
Telcordia, 243
Telecommunications Act (1996), 59, 933–934
Telephone network digital hierarchy
  See also Cable telephony
  analog-to-digital conversion, 245–247
  asynchronous, 247–249
  SONET, 249–250
  time division multiplexing, 247
  voice signal sample, 245
Telephone systems, architecture, 231
  See also Cable telephony
  backup power, 238
  call processing, 234–235
  CLASS, 239–240
  digital loop carrier, 235–238
  local exchange, 232–233
  POTS, 238
  subscriber loop, 240–244
  switch architecture, 233–235
Telephony
  service reliability and availability, 789–790, 792
  switched circuit, constant-bit-rate, 727–729
  taps, 475
  voice-grade service requirements, 727–730
Teletext, 60
Television
See also Analog television; Cable television; High-definition television (HDTV)
  broadcast digital, 72–78
  connection problems, 906–917
  direct feed, 327–330
  features of, 905–906
  integrating device, 929
  stereo systems, 368–373
Television Allocations Study Organization (TASO), 63, 64
Television Decoder Circuitry Act (1990), 59
Television Engineering Handbook (Fink), 39
Television receivers, direct connection of, 22–23
Telidon, 60, 61
Terminal equipment, 16
  examples of, 733–736
  hybrid, 736
  in-home, 734
  point-of-entry, 734–735
  shared, 735–736
  subscriber-owned, 23
Terminating taps, 474
Thermal gain control, 660
Thermal noise, 14–15, 484, 781
Thermoelectric coolers (TECs), 538
Third parties, access to networks by unaffiliated, 5–6
Threshold boundaries, 677–679
Threshold extension, 319
Through-loss, 948–949
Tilt, 462
Time division aliasing cancellation (TDAC) transform, 110
Time division multiple access (TDMA), 171–172, 257–258, 656–658, 711
Time division multiplexing (TDM), 112, 125, 172, 257–258, 710
  in telephones, 247
Time domain reflectometry (TDR), 434–436
Time domain view, laser clipping and, 683–684
Timers, 926
Time sharing, 710–711
Time slots
  assignment, 258
  contention for, 189
  reservation of, 188
Timing distribution, MPEG-2 systems transport, 119
Timing offset, 259–260
Toshiba, 890
Traffic engineering, 711, 715
Transcoding, 122
Transform coding, 84–90
Transimpedance amplifier, 543, 771
Transmission, 124
  bidirectional, 774–775
  distance limitations for analog, 780–782
  distance limitations for digital, 782–785
  forward error correction, 125–128
  loss, 501
  modulation, 125

picture quality, 128–129
signal quality, 128
slope, 572–574
Transmission Control Protocol (TCP), 224–226
Transmitters
comparison of digital and analog, 769–771
failure rates for, 804
low-frequency content removal in digital, 773–774
Transponders, 304
Transport layer, OSI, 187
Transport stream
digital program insertion, 122–124
emergency alert system, 124
MPEG-2 systems layer, 112–120
reencoding and transcoding, 122
statistical multiplexing, 120–122
Traps
negative, 20, 822–823
positive, 20, 823–825
service drop and, 500–501
Tree-and-branch topology, 14, 15
TR-08 interface, 236, 249, 250
TR-303 SONET interface, 236, 249, 250
Trellis code modulation (TCM), 180
Trivial File Transfer Protocol (TFTP), 200, 966
Trunk amplifiers, 463
Trunk controllers, 233
Trunking Gateway Control Protocol (TGCP), 269–274
Turner overload, 950
TV integrating device (TVID), 929
Twisted-pair wiring, 212–213, 253
Two-way systems, failure rates for, 809–810

Ultra directive antennas (UDAs), 311
Ultra-polished connectors, 529–530
Universal remotes, 927–928
Universal serial bus (USB), 195
Unsolicited grant services (UGS), 202
with activity detection, 202
Upstream
bandwidth, 703–704
block conversion, 653–655
common path distortion, 685
composite reverse signal, 680–681
group delay, 652–653
features, 196, 197, 199
frequency partitioning, 651–652
laser clipping, frequency domain view, 681–683
laser clipping, time domain view, 683–684
lasers, characteristics of, 668–674
lasers, specifying, 662–668
noise/interference mitigation techniques, 685–690
noise/interference, unwanted, 674–679
out-of-band channels, 854–860
rejection of signals, 950–951
scheduling types, 202
signal apportionment, examples of, 690–696
signal level management, 495–496, 655–662
two-way node, 650–651
User Datagram Protocol (UDP), 224, 269, 277
User-terminal signal processing, 717–718
Utilities, monitoring of, 730
Utility poles, access to, 10

Variable bit rate (VBR), 389–391, 711
Variable constellation multitone (VCMT), 174
VCRs
connection problems, 906–917
features of, 905–906
through-loss, 948–949
Vector quantization, 108
Venetian blinds, 67
Vertical blanking interval (VBI), 32, 58–60, 61
information carried in, 101
Vestigial sideband (VSB) filtering, 348–350
Vestigial sideband modulation, 43 8-, 153–154
V5.2 interface, 249, 250
Video
See also Digital video compression (DVC); Digital video interfaces
analog television baseband monochrome, 27–40
bandwidth, 43, 45
hum, 383–385
inversion, 834–836
switched headend structures, 747–750
waveforms, 991–996
Video distribution amplifiers (VDAs), 378
Video loop-through, 379–381
Video-on-demand (VOD), 13, 78, 388–391, 710, 839
Video services
broadcast, 718–721
narrowcast, 721–723
on-demand/interactive, 723–724
Video switch, automatic gain circuit and, 342
Visual carrier demodulation, 359–360
VOD metadata, 848
Voice activity detection, 274
Voice circuits, compression in, 273–274
Voice-grade telephony service requirements, 727–730
Voice over IP. *See* VoIP
Voice signals, sampling, 245
VoIP (voice over IP)
See also Internet protocol (IP) telephony
codecs, 273

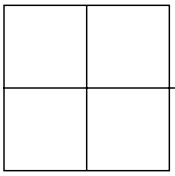
VoIP (voice over IP) (*continued*)
compression in voice circuits, 273–274
conference calling, 279–280
control protocols, 281–282
delay, 274–275
description of, 267–271
development of, 266–267
echo cancellation, 275–276
makeup of packets, 277–279
MGCP and TGCP standards, 269–274
service requirements, 729–730
signals, bandwidth requirements, 280–281
Real Time Protocol, 276–277
User Datagram Protocol, 277
voice activity detection, 274
Voltage standing wave ratio (VSWR), 633

Walk-off, 569
Water peak, 519
Waveform
characteristics of scrambled, 831
descrambled, 832–834
video, 991–996
Waveguide dispersion, 522
Wavelength
chromatic (wavelength) dispersion, 522–523
of coaxial cable, 432
cutoff, 432
dispersion, 514–516
velocity of propagation and, 510–512
Wavelength division multiplexing (WDM), 18–19, 329
amplifiers, 563–564
components for, 562–564
composite second order, 582–585
crosstalk, 565–582
defined, 559
degradation, 585–589
dense, 531
design factors, 564–565
link performance, summary of, 589–590
multiplexers, 562–563
properties of, 560–561
types of, 559–560
Wavelength multiplexers, 562–563
Wavelet, 107
WavePhore Corp., 61, 62
W curve, 65–66
White noise, 585–587
Wide wavelength division multiplexing (WWDM), 559–560
Wi-Fi Alliance, 972
WiFi protected access (WPA), 976
Windowed samples, 110
Wired equivalent privacy (WEP), 976, 977
World System Teletext (WST), 60, 61

X.509, 203, 878, 879

Yagi antennas, 293–295
Yes! Entertainment Corp., 62, 63
YPrPb interface, 863–864

Zenith Electronics Corp., 73
Zero chop pulse, 417
Zigzag scanning, 91, 92
Zworykin, Vladimir, 29, 31



About the Authors

Walter Ciciora is a Fellow of the IEEE, the SMPTE, and SCTE, and is a consultant in Cable, Consumer Electronics, and Telecommunications. He is a cofounder and CTO of HBA Matchmaker Media, a company with technologies in addressable advertising. Dr. Ciciora was cofounder and CTO of EnCamera Sciences, a company with technologies for embedding digital data in analog television signals, until it was sold in 2000. Previously, he was VP of Technology at Time Warner from 1982 to 1993 after being with Zenith since 1965.

David Large is the Chief Technical Officer of Altrio Communications. He is a Fellow Member and Hall of Fame Honoree of the SCTE, a Senior Member of the IEEE, an NCTA Science and Technology Vanguard Award Winner, and SCTE-certified Broadband Communications Engineer.

James Farmer is Chief Technical Officer at Wave7 Optics. He has previously been with Scientific-Atlanta, ESP, and ANTEC. He is a Senior Member of the IEEE and the SCTE and has served on administrative boards with both organizations. He is a recipient of the NCTA Vanguard Award in Technology, and is a member of the SCTE Hall of Fame.

Michael Adams is President of Broadband Semantics, Inc. He is a Senior Member of the IEEE, and a member of the SCTE. In 2001, he received the Cable Center book award for "OpenCable Architecture."